Privacy Preservation without Compromising Data Integrity

Tishna Sabrina

A dissertation submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

Supervisor: Professor Manzur Murshed

Gippsland School of Information Technology
Monash University, Australia

January 2014

Typeset in Palatino Linotype

Except where otherwise indicated, this thesis is my own original work and has not been submitted for any other degree.

Tishna Sabrina

January 2014

Dedicated to my mother, Begum Syeda Tahera, for all her love, inspiration, and sacrifices she made to take me up to this stage.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my supervisor Professor Manzur Murshed for his constant guidance, insightful advice, helpful criticism, valuable suggestions, commendable support, and endless patience towards the completion of this thesis. It was truly an honor to have studied under his guidance. It is he who spent painstaking hours in improving the worth and the presentation of the thesis. Without his inspiring enthusiasm and encouragement this thesis would not have been completed. I also learned many moral standards and beliefs from Professor Murshed, relating to both academic and non-academic matters.

I wish to express my gratitude to Monash University for providing an excellent environment for research and financial support. I thank all the staffs, graduate students, and friends at Gippsland School of Information Technology (GSIT), Monash University, for their support and encouragement during the last few eventful years in my life. I would also like to extend thanks to my research group members Dr Mortuza Ali, Dr. Kh. Mahmudul Alam, and Dr. Anindya Iqbal. I wish to thank my proof-reader Dr. Brendan Moloney. Special thanks to Shampa Shahriyar, SM Abdullah, Rakib Hassan, Dr. Rashidul Hasan, Dr. Kamrul Islam, Ahsan Raja Chowdhury, Dr. Shusmita Anwar Sharna, and Shaila Pervin for their kind assistance and valuable suggestions in the procedure of this thesis submission.

My heartfelt thanks go to my parents for their love, inspiration, and sacrifice. I express my special thanks to my husband Anindya Iqbal for his love, patience, understanding, and valuable suggestions during this work. I also thank my siblings Tanvir Khan, Shaila Khan and Shanta Akhter for their encouragement during my research. Special thanks to my daughter Abonti Bonhishikha Iqbal for being the constant source of refreshment and reassurance. Finally, I would like to express my profound gratitude to the Almighty for giving me the intellect and bravery to undertake this research.

Abstract

In people-centric applications, participants voluntarily report data to service providers for community benefits. As most of the applications demand high-quality data, straightforward representation of even seemingly benign data may pose significant privacy risks through inference. Retaining high data quality without compromising participants' privacy is a challenging research problem since these goals are inherently orthogonal. The existing techniques attempt to protect user privacy by reducing data precision or infusing obfuscation that ultimately degrade data quality. This thesis introduces a novel plaintext data sharing framework that aims to provide high-quality data at the desired end, protect privacy at vulnerable points such as adversaries, and safeguard against untrustworthy data manipulations. A novel subset-coding technique is developed to anonymize user reports from where original data can be retrieved through joint-decoding only if sufficient reports are received. The proposed framework is applicable when many people observe/express opinion about individual instances. Two widely-known people-centric application scenarios—participatory sensing and electronic voting—are considered. In participatory sensing, participants use data capturing devices such as smartphones that often profile their whereabouts, interests, activities, and relationships and hence, intensify inferable privacy risks. To mitigate such risks a number of anonymization and joint-decoding algorithms are proposed considering both probabilistic and deterministic decision mechanisms to cater for different participation rate e.g., commonly visited points of interests or rarely visited ones. Comprehensive adversary models are investigated and analytical privacy risk models are presented along with risk mitigation strategies. Verifiability of the received data is not of considerable significance in participatory sensing. However, wide-acceptance of electronic voting systems largely depend on guaranteeing vote-verifiability (vote is cast-as-intended) and tally-verifiability (vote is counted-as-cast) while thwarting any attempt of revealing voter-vote association to mitigate privacy, coercion, and vote-trading risks. The proposed subset-coding technique is successfully applied in this context to design an end-to-end

verifiable electronic voting framework. The strength of joint-decoding is shown robust not only to detect any vote manipulation attempts by the voting machines but also to provide individual verifiability indirectly. Different possible threats are analysed and solutions are designed accordingly. Extensive performance analysis, including computational complexity of key algorithms, are carried out with analytical models, wherever deemed possible, and rigorous simulation experiments to establish the applicability and efficacy of the proposed techniques in various realistic scenarios.

Abbreviations

AD	Autonomous Domain
ApS	Application Server
AR	Anonymized Rule
ARS	Anonymized Rules Set
AV	Anonymized Vote
BB	Bulletin Board
BGAS	Basic Greedy Anonymization Scheme
CAAS	Conforming Attribute Assignment Set
CR	Cardinality Reduction
DER	Direct Electronic Recording
DGAS	Deterministic Greedy Anonymization Scheme
E2E	End-to-End
EGAS	Enhanced Greedy Anonymization Scheme
EVM	Electronic Voting Machine
EVS	Electronic Voting System
FDGAS	Fast Deterministic Greedy Anonymization Scheme
GCA	Globally Cloaked Area
HC	Hilbert Cloak
HP^3	Hot-Potato-Privacy-Protection
IoT	Internet of Things
LAP	Lightweight Anonymity and Privacy
LCA	Locally Cloaked Area
LSH	Locality-Sensitive Hashing
MN	Mobile Node
NNC	Nearest Neighbour Cloak
OR	Observation Report
PaV	Prêt à Voter

P2P	Peer-to-Peer
PAAS	Possible Attribute Assignment Set
PAS	Possible Anonymization Subset
PA-MSN	Privacy Assurance system for Mobile Sensing Networks
PDA	Privacy-preserving Data Aggregation
PGAS	Probabilistic Greedy Anonymization Scheme
PIR	Private Information Retrieval
POIs	Points of Interest
PSS	Participatory Sensing System
RADP	Reverse Auction based Dynamic Price
RNG	Random Number Generator
TrPF	Trajectory Privacy-preserving Framework
TPM	Trusted Platform Module
VPC	Virtual Participation Credit
WSNs	Wireless Sensor Networks

List of Publications from this Research

Papers that have been published from this research

1. T. Sabrina, and M. Murshed, "Privacy in Participatory Sensing Systems," In J. Abawajy, M. Pathan, M. Rahman, A. Pathan, & M. Deris (Eds.), Network and Traffic Engineering in Emerging Distributed Computing Applications (pp. 124-143).
2. M. Murshed, T. Sabrina, A. Iqbal, and K. H. Alam, "A novel anonymization technique to trade-off location privacy and data integrity in participatory sensing systems," Proc. of IEEE Intl. Conf. on Network and System Security (NSS), 2010.
3. M. Murshed, A. Iqbal, T. Sabrina, and K. H. Alam, "A subset coding based k-anonymization technique to trade-off location privacy and data integrity in participatory sensing systems," Proc. of IEEE Intl. Symp. on Network Computing and Applications (NCA), 2011.
4. T. Sabrina and M. Murshed, "Analysis of Location Privacy Risk in a Plain-text Communication Based Participatory Sensing System Using Subset Coding and Mix Network," International Symposium on Communications and Information Technologies (ISCIT), 2012.
5. M. Murshed, T. Sabrina, A. Iqbal, and M. Ali, "Verifiable and Privacy Preserving Electronic Voting with Untrusted Machines," IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2013.

Papers in Preparation

6. T. Sabrina and M. Murshed, "Deterministic Greedy Anonymization Schemes for simultaneous privacy preservation and maintaining data integrity," In preparation for *IEEE Transactions on Information Forensics* and *Security*.
7. M. Murshed, T. Sabrina, A. Iqbal, and M. Ali, "Verifiable and Privacy Preserving Voting with Minimal Trust Components," In preparation for *IEEE Transactions on Information Forensics* and *Security*.

Table of Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

I'm Nobody! Who are you?
Are you – Nobody – too?
Then there's a pair of us?
Don't tell! they'd advertise – you know!

Emily Dickinson

1.1 Introduction

In the current era of rapidly advancing technology, people are becoming accustomed to the service provided by it, but not always understanding the privacy or trust issues involved. In general, **privacy** can be defined as a preference for maintaining the confidentiality of personal information. On the one hand, the increasing popularity of online shopping sites like eBay or social networking sites like Facebook or Twitter, support the idea that we are entering into a whole new age of a virtual world. On the other hand, fraudster and hackers are also taking advantage of weaknesses in modern technology and becoming more innovative and sophisticated in their privacy and security attacks. The arrival of wearable technology and the evolution of the Internet of Things (IoT) which combines inexpensive, remote sensors with Big Data analytics have the potential to threaten and undermine long-held concepts like personal privacy and the rights of individuals. Consequently, our ever-increasing dependency on technology for daily activities may involve many vulnerable exposures to fraud. Recent growing incidents of credit card fraud, the impersonation of

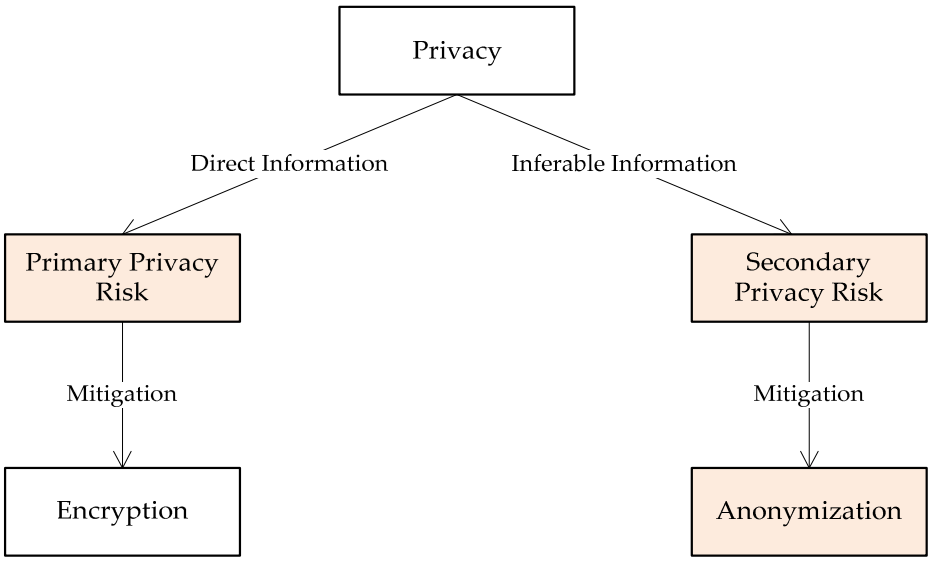


Figure 1.1: Classification of privacy risks with mitigation policies.

Facebook accounts and even the latest launching of the classified media WikiLeaks have impelled people to focus on the significance of privacy.

In the digital world of communication, privacy risks arise from intercepting information as it flows in the network. The highlighted boxes in Figure 1.1 refer to the areas we aim to examine in this thesis. As shown in Figure 1.1, there are mainly two types of privacy risks, named as primary and secondary privacy risks. The first one is the risk evolving from direct interception of personal information that the individual prefers to keep private. In this case, encryption can be used to mitigate the risk of breaching primary privacy. On the one hand, in the field of digital communication, there has been a huge chunk of research done on direct data encryption. The second type of privacy risk, on the other hand, involves the seemingly benign information, e.g., personal whereabouts that may in turn deduce direct information and, thus, breach primary privacy. From the concept of Frequency Attack, we know that the most highly probable data and the most frequent data in the encrypted domain are most likely to be the same. Hence, probabilistic analysis from this apparently harmless information may help the frequency attack to break encryption. For example, when a particular person is visiting a petrol pump near a cancer hospital most frequently, then an adversary with some more prior information may assume that person to be a cancer patient. This implies that secondary inferable information is also necessary to be protected. However, encrypting all communications is neither desirable nor applicable. In many application scenarios, we need to share the benign information in the public domain. For

that purpose, we prefer anonymization to mitigate secondary privacy risks. The aim of anonymization is to introduce uncertainty in the information conveyed such that getting actual information is no longer straightforward. There have been very few works on secondary privacy risks and the field is still immature in comparison to work done on primary privacy risks.

While the privacy requirement of users needed to be ensured, data integrity in contrast is also undeniably desirable in making service dependable and comprehensible. In general, **data integrity** refers to maintaining and assuring the accuracy and consistency of data over its entire life-cycle and is a critical aspect to the design, implementation, and usage of any system which stores, processes or retrieves data. Even a very simple service like sending a document electronically may involve several data service quality issues, e.g., whether it was actually sent from the sender as shown or not, whether it was changed in its way to the receiver or not, whether there was any transmission delay or not, and so on. It is notable that a top priority of every user is an error free data communication service. A substantial amount of research work has been done in this area of data communication. However, it is also true that data integrity and user privacy are somehow orthogonal. Database literature also supports this concept in the Bell-La Padula model and Biba model [1] where the former ensures security at the expense of integrity and the latter ensures integrity at the expense of security. The Bell-La Padula model is characterized by the phrase "no write down, no read up", which ensures that data secrecy comes at the expense of data integrity. In contrast, the Biba model is characterized by the phrase "no write down, no read up", which ensures data integrity, but only at the expense of compromised secrecy. When the user id is kept private, none can ensure verifiability of the data sent. Therefore, the intended receiver cannot get the actual information of the whole picture. Hence, there should be a wise trade-off of data integrity and user privacy depending on the application of that particular technology and

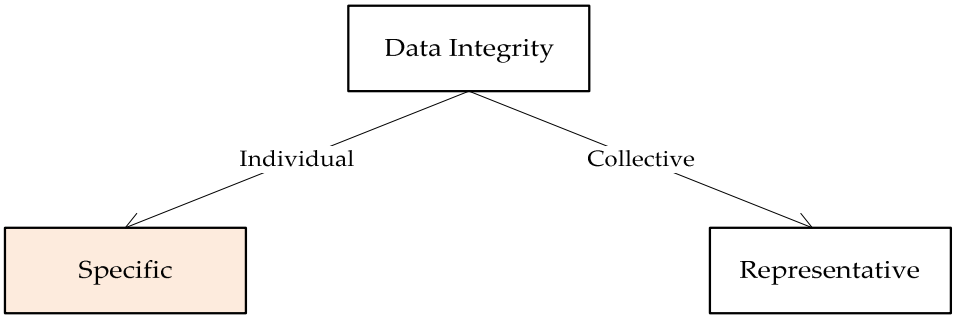


Figure 1.2: Classification of data integrity.

the user's requirements. Thus far, no solution is developed yet that addresses this universal problem. We aim to work with this problem in a particular context.

In general, there can be two different illustrations of data integrity, namely specific and representative data quality as shown in Figure 1.2. The Specific box denotes individual data integrity where data on individual entities are collected and are intended to be retained. In this case, data quality is measured by specificity, i.e., how specifically the system is able to retain that data. In contrast, collective data quality means the aggregated data which should correctly represent the collection, e.g., average salary in a profession. In most of the cases, the collective data quality is measured by the expected or mean value of the collection. Anonymization in general can significantly compromise individual data quality. However, some anonymization techniques exist, such as adding random Gaussian noise with a zero mean that can preserve the expected value of the collection. Hence, we are mainly interested in developing novel anonymization techniques that can preserve individual data quality, while applying anonymization for privacy risk mitigation (as highlighted in Figure 1.2). When information is shared in a system and there is a probability to obtain the same information from multiple observers, then we can hypothesize that preserving privacy without compromising data quality may be possible by exploiting redundancies in multiple observations. In this thesis, we attempt to establish this hypothesis. For this purpose, we have selected a particular system approach, **people-centric application** that upholds the property of information collection where multiple entities may report the same information.

Realizing the infinite potentiality of mass contribution, many people-centric applications have evolved to provide the common people a platform to participate and to achieve goals that traditional organizations have failed to do. However, in this scenario, the success of the system is very much dependent on the trustworthiness of the contributors. Hence, trustworthiness is another aspect to consider. Here, **trustworthiness** is the ability to detect the fraudulent behaviour of the sender with a high probability. The approach of most existing privacy preserving techniques is to add some uncertainty in individual observation/choice which often stands in the way of maintaining quality/integrity at the desired end. This thesis aims to design an intelligent privacy preserving scheme such that the desired properties are made prominent at the desired ends.

1.2 Significance

The significance of the contexts we are referring to is discussed here to understand the relevance of the problems and also the reason why these have attracted research interest in recent time.

Recently many digital application scenarios like Participatory Sensing System (PSS) are emerging where community people share the apparently insensitive information that travels across open wireless networks. The advancement of wireless communication technologies has facilitated the development and popularity of mobile devices equipped with powerful sensing, storage, and processing capabilities. Unlike web applications, data is sensed using ad-hoc sensing devices mounted on, for instance, cell phones, and vehicles from Points of Interest (POIs) that the participating users visit in the course of their daily life. These are then sent to servers via lightweight and inexpensive wireless communication networks. Target servers aggregate data received from users and reply to the queries accordingly. PSS has a wide range of real-world applications including consumer price sharing [2]-[5], measuring safety in localities [6][6], variation in elevation along bike routes monitoring [7], vehicular transportation monitoring [8]-[10], public health such as monitoring the effectiveness of diet programs [11], environmental impact and exposure [12], urban planning [13], sound events [14] , earthquakes [15], parking availabilities [16], comfort management of building [17], and prediction of bus arrival time [18].

Considering the vast range of applications, the role of PSS is no longer limited to being a mere communication medium. Rather it has become a major tool to bridge the gap between data feed from the sensing devices and human information requirements. The huge popularity of social networks e.g., Facebook and Twitter, and the existence of numerous blogs, where contributions from general people develop the content, clearly indicates the potential of PSS. WikiLeaks, identified by many as an example of participatory journalism has initiated a new era in supplementing traditional media. A similar significant role is expected from participatory sensing as a supplement to a limited number of traditional sensor networks. PSS is emerging as a cost-effective alternative for reliable and impartial data collection, processing, and dissemination. It provides a framework to facilitate communities to sense, collect, analyse, and share local information or knowledge for mutual benefit. However, handling such seemingly benign information may cause a secondary privacy risk. In this study, we aim to mitigate this issue.

To understand fully the extent of the location privacy risk via participatory sensing we need to consider the bigger picture. Most of the smart-phones are now equipped with a high precision localization capability, which can potentially leave a long trail as to their whereabouts. Many applications running on these devices also exploit this capability to offer so-called location-aware services that eventually profile an individual's habits, interests, activities, and relationships. The whereabouts of the participants may be inferred by an *adversary* if some reported data is captured by eavesdropping. While the sensed data itself may be considered insensitive/benign where privacy is concerned, the same may not be true for inferable information such as the whereabouts of participants. This is formally termed as *location privacy*. With some prior information, knowledge of location may also compromise inferable privacy. Reporting nearby a specialized medical treatment facility may assist in speculating on a reporter's medical condition if someone knows only the time of her doctor's appointment. Apart from location, an association with a product may sometimes cause privacy concerns. In consumer price sharing from a super store, for example, it is necessary to include the name of the product even when this is not desirable in some sensitive cases. People may prefer to keep the purchase of a certain drug, drink, or cigarette, for example, a secret. Protecting participating users' privacy is the primary requirement to make PSS popular. The introduction of the Location Privacy Protection Act of 2011 bill in the US Senate [20][20], in light of the Electronic Communications Privacy Act, and the Video Privacy Protection Act, clearly emphasises the level of risk involved. We know that the success of WikiLeaks is directly linked to objective and ability to keep sources anonymous. Similar factors will prevail in the success of PSS.

While ensuring privacy, the integrity of data also needs to be maintained to ensure reliable response to queries. For example, in considering a consumer fuel price sharing system in [1][2], if the fuel price at the suggested fuel station is not the cheapest, due to loss of data quality, dissatisfied users will not participate in future. Ultimately, the best case scenario is in achieving anonymity at an individual (adversary) end and data integrity at the service provider end. Because the ultimate service can be provided only when the service provider can interpret reported data at a much higher accuracy. Data integrity is undoubtedly orthogonal to security/privacy. It is supported by the assumption that no one can maintain privacy from the entity which is supposed to provide full data integrity. Hence, achieving an acceptable level of an observer's privacy and simultaneously maintaining data integrity is crucial to satisfactorily ensuring the voluntary participation of

a critical mass. This privacy-data integrity trade-off model should be such that either the privacy requirement will dictate achievable data integrity level or vice versa.

An important contribution from 'the people' in the modern age is the democratic election of their representative. Voting, or polling, is the fundamental requirement of democracy a system which is believed to be the best political system under the utilitarian assumption that the majority would not make wrong judgement. Designing an Electronic Voting System (EVS) is a natural goal to enjoy the benefit of digitization. An EVS can be user-friendly, and provide fast and flawless counting, auditability, and verifiability. In modern societies, while every possible task is done electronically to maximise efficiency, reliability, and accuracy, voting in most countries is still based on paper and ballot boxes. This can be attributed to the simultaneous technical challenges in achieving all the requirements of an effective and reliable voting system. Besides meeting the requirements common to any technology based system, an EVS must address the requirements unique to voting. The same is true for online surveys, which are gaining popularity day by day. According to ESOMAR, online survey research accounts for 20% of global data-collection expenditure in 2006 [19][19]. People must be assured of their identity privacy preservation so that they can express their opinions without any reservation.

Achieving both privacy and verifiability/trustworthiness in an electronic voting system is challenging due to the inherent properties of the electronic machines. Electronic Voting Machines (EVMs) are, in essence, general purpose computers consisting of sophisticated hardware and software. They essentially consist of a processor, memory, input/output interfaces, and an operating system. Clearly, the behaviour of the EVM at a particular instance depends on the installed software/firmware. Besides, the machine must allow either removable media or communication capability to convey the results to the central election authority for processing. All these components are vulnerable to attack both from inside and outside of the election authority. Thus, trusted EVM is a hard assumption in practice.

These two apparently dissimilar contexts have many features in common, particularly in the sense that many people observe or make choices from a pre-defined set of objects. In the context of PSS, the objects refer to various POIs or simply objects/products. In voting/surveys, choices have to be made from a set of candidates/survey answers. Another similarity is in the requirement of success: the more people participate, the better. Again, participation of the people relies a great deal on the assurance of their privacy. Hence,

privacy or anonymity of the participants is a major concern in both these contexts. For PSS, data quality at the destination provides integrity to the system. In the case of voting/survey, verifiability that the vote/feedback is actually counted or free of any manipulation brings integrity and trustworthiness to the system. This thesis presents a novel *subset-coding* technique that achieves both privacy and integrity in both these contexts. This technique is applicable to many applications that have the property of "multiple observations of independent instances."

1.3 Motivation

To mitigate the risk of privacy, a number of location privacy protection mechanisms have been proposed using the techniques of *k*-anonymity, obfuscation, mix-zones, or dummy locations [21][21] (elaborated on in Chapter 2). The underlying principle of all these mechanisms is to record location information with some anonymity or by adding Gaussian noise or with reduced precision so that any probabilistic attempt to decode the exact location or track remains ambiguous. The safeguard offered by such ambiguity can be severely compromised if the exact locations of an individual are known at some temporal points. The standard mechanisms moreover, cannot be used directly where the destination expects

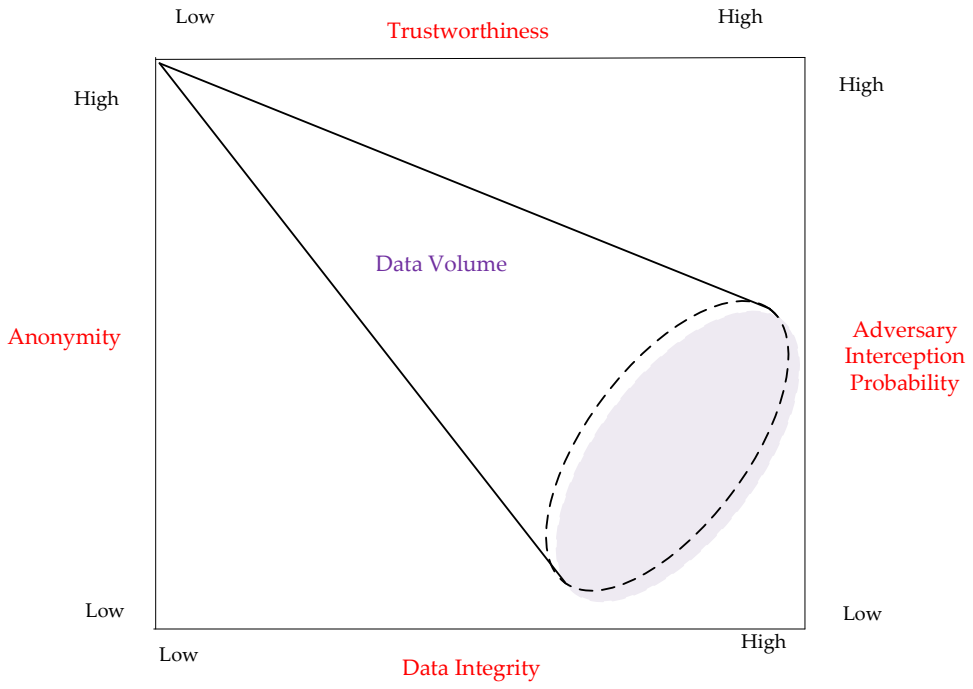


Figure 1.3: Basic idea of proposed approach.

complete data integrity at an individual level e.g., the *PetrolWatch* [2] that assists drivers to find the cheapest fuel station in the neighbourhood. Such PSSs need a privacy-preserving data communication technique so that each observation from a participant can be transmitted with sufficient anonymity such that the data collector is able to de-anonymize/decode individual data only through the joint processing of the entire collection. So long as an adversary is unable to intercept a reasonably high number of transmissions from the participants, any de-anonymization attempt to infer sensitive information remains sufficiently ambiguous.

Figure 1.3 describes our basic approach to deal with data integrity, anonymity, trustworthiness/reliability, and adversary interception probability. It uses four axes for four of the major issues in a basic communication system; the dataflow is shown in volume. In digital communication, an adversary is defined as the entity who can intercept a digital transmission. When data volume is low, it in turn guarantees a high anonymity in spite of high interception probability and low reliability. Trustworthiness or reliability is indeed parallel to data integrity. We may worry less about the trustworthiness/reliability issue where data integrity is guaranteed. In PSS, our main target is to provide specific data integrity at the service provider's end. Trustworthiness/reliability is maintained with high data volume because our joint decoding scheme should be able to detect false data feeding. Hence when we aim to achieve the main target, i.e., the specific data integrity, we can achieve the trustworthiness inherently. This is how they are related to data volume and can successfully facilitate the service provider. In contrast, the adversary intercepts digital communication and works with this low data volume. Here, our aim is to achieve high anonymity accompanied by low data integrity. Hence we can confirm that all the desirable properties can be managed intelligently in our approach such that the required property is maintained at the relevant point. Note that privacy by secrecy cannot ensure trustworthiness. This is another reason why encryption cannot be used here.

In PSS, there are many senders, whereas in EVS, the only sender is the voting machine that collects the votes and sends them to the bulletin board. When there are many senders, the system does not worry much about the accountability or verifiability. In general, **verifiability** can be defined as the ability to ensure that the receiver is using the same data as the sender originally sent. In PSS so long as the service (i.e., the reply to query) is correct, the system does not bother with verifiability. For example, in the case of *PetrolWatch* [2] even the

second cheapest petrol pump is acceptable and the participants have no interest bearing extra communication to verify their observations being recorded as reported. However, in voting we require data integrity as an absolute certainty, that is, any sort of vote manipulation attempt is not acceptable. Hence, verifiability appears as another cornerstone of an efficient EVS. Individual data in consideration of its contribution in PSS is important but certainly not as important as the individual vote.

The primary requirement of any voting system is confidence or trust among voters as to the outcome of an election. This goes beyond the classic security properties of a system such as confidentiality, integrity, and availability. Verifiability and auditability are two natural responses to this demand in any electronic system. A simple way of maintaining verifiability is to provide a receipt (copy) of the vote to each voter which she can verify from a bulletin board later on. After the voting period, all votes would be displayed on the bulletin board in anonymous, yet an identifiable, form. Both the election authority and a third party auditor would count votes using these. However, this approach conflicts with other mandatory requirements such as a voter's privacy, and resistance to vote trading and coercion. A voters' privacy must be protected by establishing the un-linkability between a voter and the vote she casts. From a privacy point of view, it is the primary privacy right of the voter to maintain this un-linkability. From Figure 1.1, we know that the primary privacy risk can be mitigated using encryption. However, encryption is not sufficient to guarantee trustworthiness which is a primary requirement here. That is why this second scenario is considered where encryption is not enough to address even primary privacy risk. The same approach that we propose for mitigating secondary privacy risk in PSS can be used intuitively/innovatively for mitigation of the primary privacy risk in EVS by using anonymization. To prove our hypothesis we need to address both these two scenarios of people centric application. Vote manipulation attempts at any machine or communication medium involved in the whole process of EVS from vote cast to counting has to be checked. There is also the risk of false accusation of fraud by a losing party. It has to be ensured that a false allegation cannot be proved by any means. All these indispensable requirements have made electronic voting a significantly challenging research issue.

Electronic voting systems have begun to replace the traditional paper ballot based system in the United States since the 1980's. However, the vulnerability of the EVMs to security threats came to public attention in the wake of the 2000 US Presidential

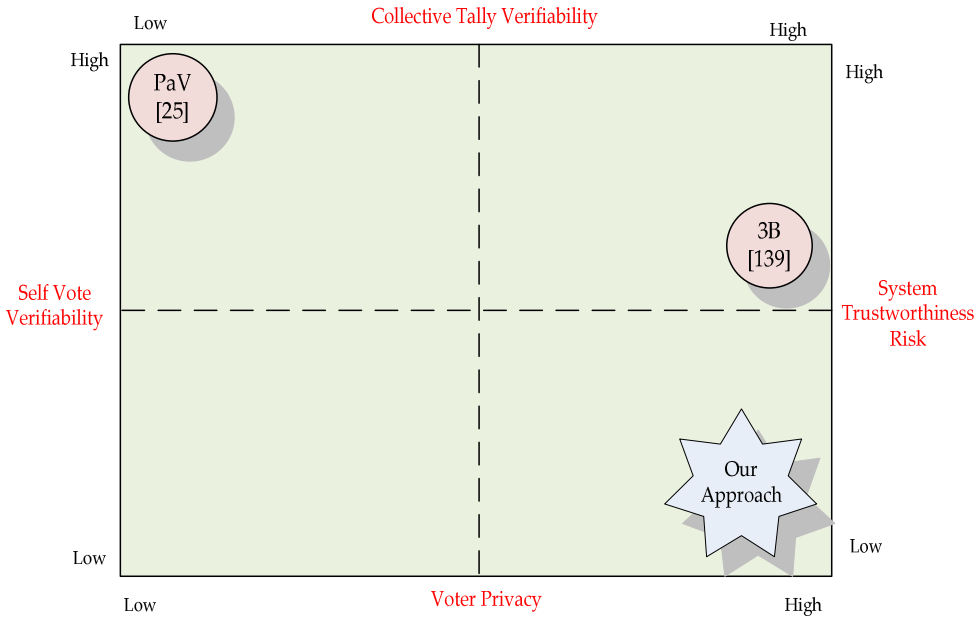


Figure 1.4: Research challenges in EVS as dealt with in contemporary approaches.

Election [22]. Therefore, a recent trend is to replace EVMs with a paper ballot system. It is noteworthy that the state-of-the-art voting systems proposed in the literature, such as Scantegrity [23], Scantegrity II [24], Prêt à Voter (PaV) [25], Split-Ballot [26], ThreeBallot (3B) [27], and Hover [28], use pre-printed paper ballot. In essence, the paper ballot in these schemes is scanned just after casting and sent to the central election authority for the efficiency of counting and processing. The scanned ballots are also posted to a bulletin board for verifiability. However, the optical scanners, with communication and/or storage capabilities, are also subject to integrity vulnerability.

Figure 1.4 shows our basic approach to deal with voter privacy, self-vote verifiability, collective tally verifiability, and system trustworthiness risk. Our proposed approach aims at high voter privacy, low risk on system trustworthiness, and high tally verifiability. However, it suffers only from self-vote verifiability. The consequence of ensuring verifiability is that often either privacy or trustworthiness has to be compromised. That is why, in our approach, we aim to ensure verifiability indirectly such that these consequences can be avoided. To mitigate it we take help of *joint decoding* which states that any attempt of manipulation will be detected by bringing inconsistency in joint decoding. The comparative superiority of our approach against other contemporary approaches will be evident when we discuss them in Chapter 4.

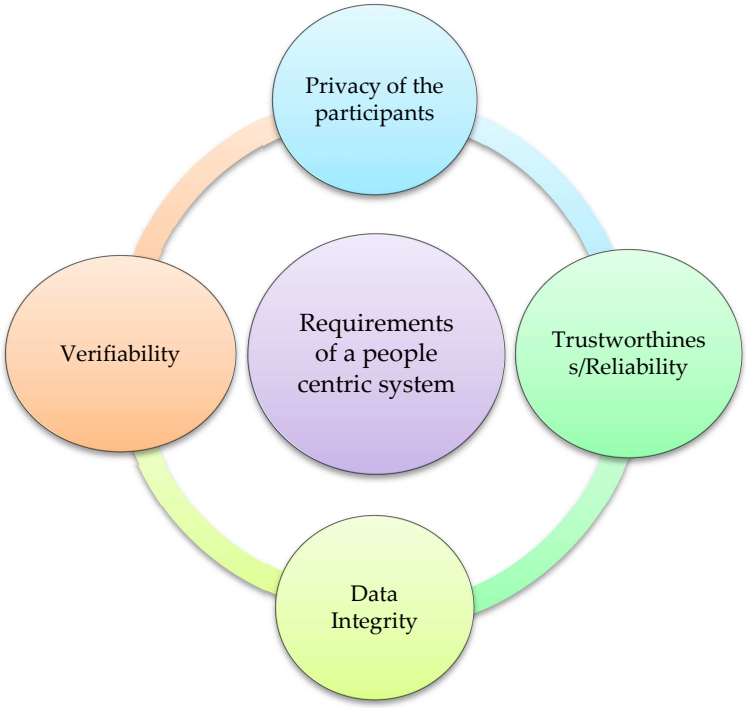


Figure 1.5: Research challenges in people centric applications.

Finally, for both these contexts, the requirements of privacy and data integrity need to be satisfied in an efficient and cost effective way. The system otherwise would not be widely acceptable. A major component of achieving efficiency is to design the anonymization/de-anonymization techniques in a computationally efficient manner. This additional requirement makes the problem even more challenging. In detail, there are actually four cornerstones of any people centric applications as depicted in Figure 1.5. Verifiability is not much highly valued in PSS. That is why we also work with another application scenario, EVS, where verifiability is a must.

1.4 Aims

The aim of this thesis is to contribute to the development of a framework that simultaneously achieves the goal of preserving privacy for reporting users and maintaining integrity at the desired end. The solution is expected to address all the four issues in Figure 1.5 as well as the associated concerns and challenges stated above. This thesis, therefore, focuses on achieving the following specific aims:

1. To develop a coding based theoretical framework that anonymizes multiple individual observations of a single instance in such a way that sufficient integrity is

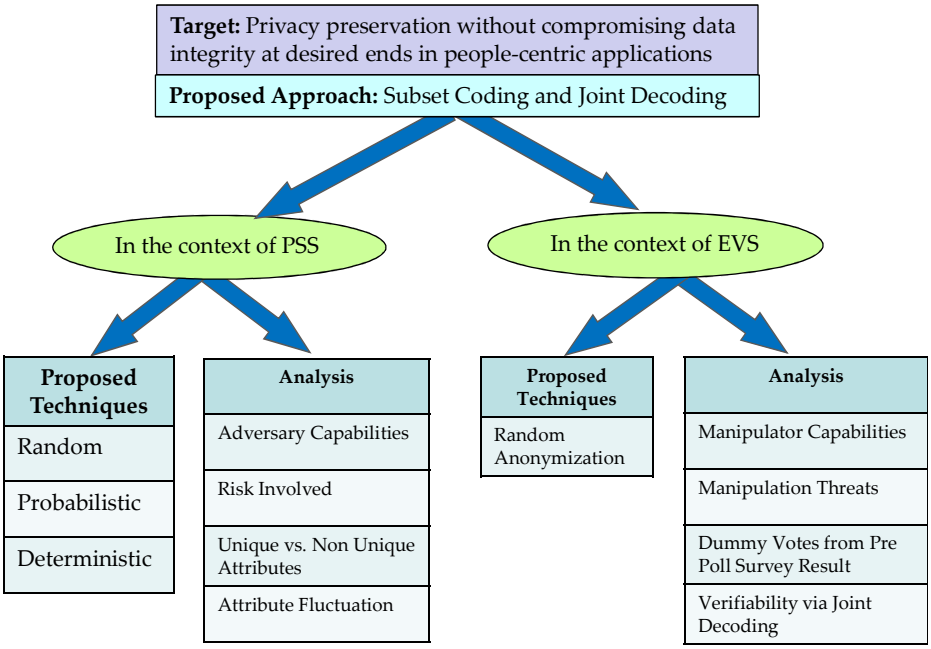


Figure 1.6: Conceptual flow diagram of this research.

achieved when observations are de-anonymized through joint information processing at the destination.

2. To design infrastructure for inferable privacy-preserving PSS and thoroughly analyse the privacy risks against different possible adversary attacks when the proposed anonymization technique is used.

3. To investigate various optimization and implementation issues of the proposed anonymization technique in the context of PSS.

4. To design an EVS that uses the same coding framework to establish the privacy of voters.

5. To analyse the trustworthiness of the EVS by investigating the scope of vote manipulation and measuring the system's ability to detect manipulation attempts.

1.5 Contributions

Figure 1.6 presents a conceptual flow diagram of the key contributions of this thesis. To conclude this introductory chapter we identify and summarise the main contributions of the thesis below:

1. Devising a novel subset-coding technique that protects the privacy of the participants by introducing anonymity **(Chapter 3)**. This contribution achieves our aim of maintaining privacy to sufficiently confuse an adversary.
2. Devising a novel joint-decoding technique to bothachieve high data quality at the desired end and to detect false data feeding **(Chapter 3)**. Using this technique, we complete our aim of avoiding sacrificing data integrity while maintaining user privacy.
3. Designing a system architecture to guard against eavesdropping in an insecure communication channel **(Chapter 3)**. This contribution achieves our aim of developing a robust system design to address all possible risks.
4. Presenting a comprehensive adversary model which identifies different malicious capabilities in the context of PSS and designing strategies to mitigate attacks on the privacy of participants **(Chapter 3)**. Here we attain the aim of addressing potential adverse threats while maintaining public network dataflow plain-text.
5. Presenting a detail risk analysis on the privacy of observers when our proposed subset-coding technique is used **(Chapter 3)**. This contribution accomplishes our aim of thoroughly analysing all possible privacy risks and keeping them within a user defined threshold.
6. Devising a greedy algorithm that *k*-anonymizes participating users in terms of location using a probabilistic subset coding scheme that aims to maximize integrity of reported data at the destination. The algorithm flexibly accommodates any value of *k* **(Chapter 4)**. Using this algorithm, we conceive of a basic approach to deal with the issue of simultaneously maintaining privacy and data integrity.
7. Enhancing our greedy anonymization algorithm by considering a number of alternative heuristics to achieve almost lossless data integrity **(Chapter 4)**. This contribution achieves our aim of optimizing our basic approach.

8. Analysing and implementing a randomized variant of the greedy approach with a comparative performance study **(Chapter 4)**. This addresses our goal of undertaking comparative performance analysis.

9. Analysing the impact of transient changes in the attributes of POIs **(Chapter 4 and 5)**. This contribution meets our aim of making our approach more realistic when facing practical issues.

10. Devising a greedy algorithm that *k*-anonymizes participating users in terms of location using the subset coding scheme that aims to obtain deterministically full data integrity of reported data at the destination. The algorithm flexibly accommodates any value of *k* and for the highest possible value of *k*, a computationally faster variant of this deterministic method is devised **(Chapter 5)**. We accomplish the target of making our scheme computationally less complex.

11. Designing a privacy preserving and verifiable voting scheme with a minimal trust component where any attempt of manipulation can be easily detected **(Chapter 6)**. This contribution meets our aim of using the same coding framework to protect the privacy of voters.

12. Analysing and investigating vote manipulation and measuring the ability of EVS to detect manipulation attempts **(Chapter 6)**.

1.6 Organisation

The organisation of the rest of the thesis is as follows.

Chapter 2: Background and Related Works This chapter defines the important terms, algorithms, and models for PSS and EVS. First, a short background of the research area is provided, including basic concepts and system architecture. As we propose to work with the privacy issues in PSS, this chapter includes a brief taxonomy of privacy preserving approaches found in the related research works. Finally, we also analyse the existing e-voting methodologies as found in the literature to investigate the privacy and verifiability issues in EVS. A publication from this chapter as a book chapter in [31].

Chapter 3: Subset Coding and Joint Decoding with Risk Analysis in PSS Here the proposed PSS system architecture with a novel subset-coding and joint-decoding based technique is introduced, and explores potential adversary attacks and provides a

comprehensive analysis on the risks on privacy of the participants. Parts of this chapter has been published in International Symposium on Communications and Information Technologies (ISCIT), 2012 [32].

Chapter 4: Probabilistic Greedy Anonymization Techniques to Achieve Location Privacy and Data Quality in PSS This chapter presents a subset-coding based probabilistic *k*-anonymization technique. Findings from this chapter have been published in IEEE International Conference on Network and System Security (NSS), 2010 [33] and IEEE International Symposium on Network Computing and Applications (NCA), 2011 [34].

Chapter 5: Deterministic Greedy Anonymization Techniques to Achieve Location Privacy and Data Quality in PSS A subset-coding based deterministic *k*-anonymization technique is illustrated in this chapter. Findings from this chapter is to be submitted in IEEE Transactions on Information Forensics and Security.

Chapter 6: Verifiable and Privacy Preserving Electronic Voting with Untrusted Machines This chapter introduces a trustworthy electronic voting system where the privacy of the voters is protected using subset-coding based *k*-anonymization. This research work has been published in IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom) 2013 [35].

Chapter 7: Conclusions and Future Works Finally, this chapter concludes the thesis and presents discussions on future research direction.

Chapter 2

2 Background and Related Works

The objective of this chapter is to define frequently used terms, introduce commonly used system entities, explain important existing algorithms, and depict different models for inferable privacy preservation while maintaining service quality in the context of people centric applications. This chapter critically reviews the literature in related contemporary research areas. We present separately the background and related works on the two application contexts, i.e. participatory sensing and electronic voting. For the first context, relevant research on both location and data privacy problems are discussed along with a brief overview of associated issues. For the second context, the works that address the problems of simultaneous privacy and verifiability are reviewed.

This chapter is organized as follows. In Section 2.1, the issue of privacy preservation in participatory sensing is introduced with a brief background, followed by reviews of relevant works that deal with the problems of location privacy and data privacy. Section 2.2 presents the background of EVS followed by some discussion on research on verifiable and privacy preserving voting. Finally, in Section 2.3 the drawback or limitations of the existing literature in both contexts are explored.

2.1 Privacy in Participatory Sensing

Participatory sensing facilitates a system providing cost-effective, reliable, and impartial data collection, processing and transmission. However, in practical terms, no one would be generous enough to contribute voluntarily if their privacy is not protected. The right against unsanctioned invasion of privacy by the government, corporations or individuals is part of many countries' privacy laws, and in some cases, constitutions. Working within the scope of privacy laws and meeting specific privacy requirements of the contributor is a must to run

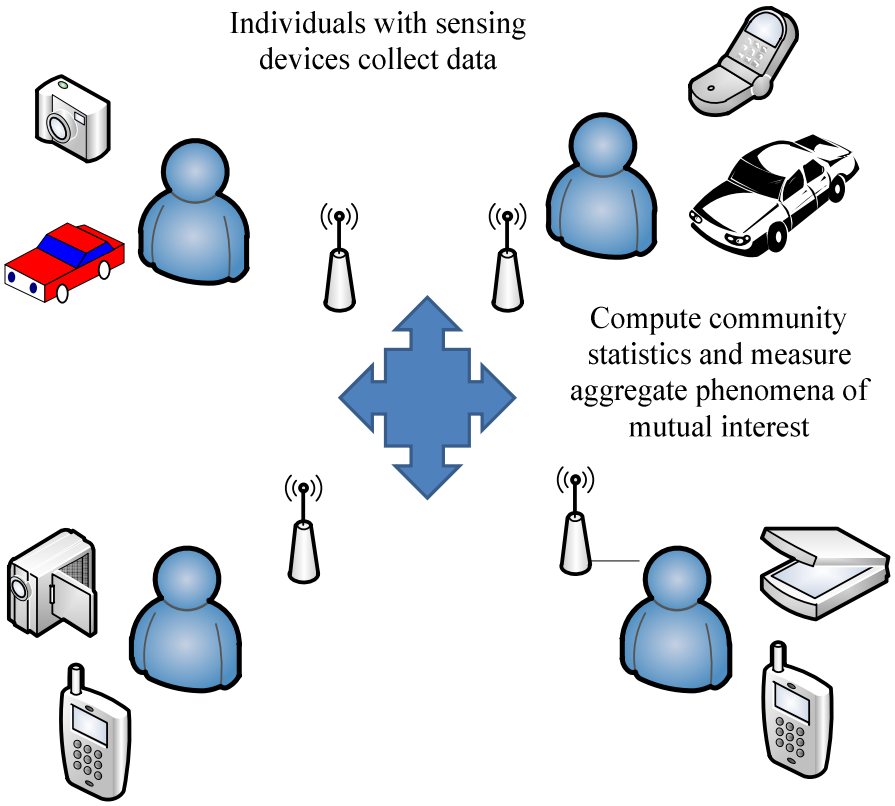


Figure 2.1: Basic outline of a participatory sensing system.

the system effectively. At the same time, in order to keep the service dependable and attractive to users, the data of interest should be credible and the quality of service should meet users' need.

This section introduces and reviews the system of participatory sensing as an emerging system with the intrinsic challenges of meeting privacy requirements and maintaining data integrity. It also presents a comparative study of various solutions offered in the literature so far. It highlights various privacy issues in PSS and examines possible approaches to face privacy attacks while at the same time maintaining data integrity. Then we will also examine pros and cons of various existing privacy preserving approaches. Among the approaches some are computationally less expensive and real-time in operation while others may be more applicable in practical scenarios. Some concentrate on preserving privacy regardless of the cost of compromising the data integrity, while some overcome the dependency on a centrally trusted node.

In Section 2.1.1 we provide a short introduction to privacy preserving approaches in participatory sensing. Section 2.1.2 presents the works that deal with the problem of location privacy. A study of data privacy problem is presented in Section 2.1.3. In Section 2.1.4, a brief overview of some related issues and how they are addressed is also provided.

2.1.1 Background

The concept of PSS [38], [39] was proposed a few years ago as a system that facilitates a community sharing data for mutual benefit. In some studies, PSS is referred to as *urban sensing* [39], or *people-centric sensing* [40] or sometimes *opportunistic sensing* [40]-[41]. Its common characteristic is that it is initiated by ordinary citizens using their privately-owned sensor-equipped mobile devices to collectively measure and share information of mutual interest from their environment. The concept has become very popular lately with the massive boost in usage of mobile devices capable of capturing, classifying and transmitting images, sounds, locations and other data, interactively or autonomously [38]. Unlike web applications, data is likely to be sensed from different places people visit in the course of their daily life using ad hoc sensing devices mounted, for example, on cell phones and vehicles. The data is then sent to servers via inexpensive wireless communication architecture. The server is able to generate aggregate results using the data received from all participating users. Using that knowledge, it then replies to the queries made by the users at any time. In short, it is a system by the people and for the people.

With participatory sensing, the mobile user consciously opts to meet an application request out of personal or financial interest. A participatory approach involves people into significant decision stages of the sensing system such as deciding what data is shared and to what extent privacy mechanisms should be allowed to impact data fidelity. In its common variant opportunistic sensing, the mobile node may not be aware of active applications. Instead, a mobile device is utilized whenever its state matches the requirements of an application. This state is automatically detected; the owner of the device does not knowingly change the device state for the purpose of meeting the application request.

The very basic architecture of a participatory sensing network consists of a collection of Mobile Nodes (MNs), some Points of Interest (POIs), and an Application Server (ApS) which is, most of the time, found to be a location-based service provider. The individual MNs that constitute mobile sensing infrastructure are devices with sensing, computation, data storage, and wireless communication capabilities. These MNs are mostly carried by humans or

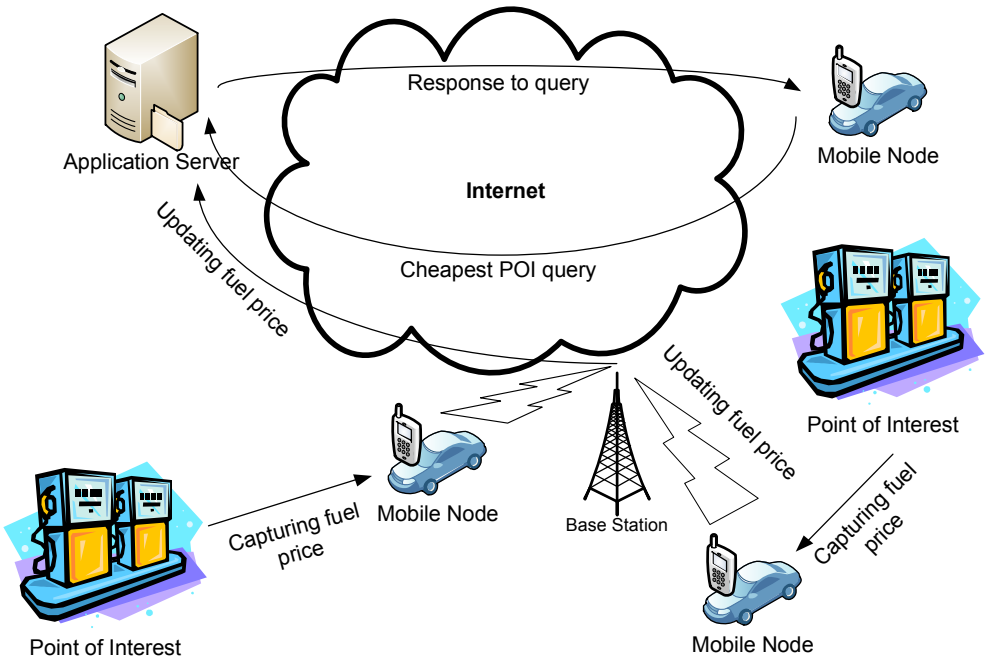


Figure 2.2: Basic architecture of a participatory sensing network *PetrolWatch*[2].

attached to other moving objects such as vehicles. The POIs are the objects whose specific attribute information is to be captured by the MNs. The MNs collect and report the particular attributes of the POIs to the ApS. The ApS is the server that receives reports from MNs and, based on these reports; it makes the service available for the user e.g., informs users about the price of petrol pumps in their vicinity. The ApS is tasked to provide the attribute information on demand from the users.

A very common application of a participatory sensing system is *PetrolWatch* [2]. In that participatory sensing application, users automatically collect, contribute, and share fuel pricing information using camera enabled mobile phones mounted on the car dashboard. Whenever the vehicle approaches a service station, the *PetrolWatch* recognizes the position through the use of GPS and GIS and the camera is automatically triggered to take snaps of fuel pricing billboards. These pictures are processed by computer vision algorithms to extract fuel prices. The fuel prices are annotated with location coordinates of the service station and the time at which the capture took place and then the whole information is sent to the ApS. Users can query the ApS to locate the cheapest petrol station in their vicinity. The ApS responds to the query based on the database developed with the data contributed.

From the basic application scenario discussed above, it is obvious that PSS has some fundamental requirements. There are some issues regarding a user's location privacy as privacy must be protected to bring this concept into reality. Who will risk her location privacy to report messages for the benefit of community? Because, we consider contributors to provide their identifications while reporting, this is a need to facilitate the development of a reputation scheme that is required for some applications. Moreover, this will allow the system developing a reward-based service provision e.g., the more someone contributes the higher will be the quality of data received from the server. Consequently, reports to the server cannot be anonymous. Hence, hiding the data ownership straight-away cannot provide a complete solution in this context. Hardly anyone, however, will be willing to report from controversial places e.g., a casino in a conservative society. With some prior information, knowledge of location may also compromise inferable privacy, for example, reporting nearby a specialized medical treatment facility may assist in speculating on a reporter's medical condition if someone knows the time of her doctor's appointment. The success of WikiLeaks is directly linked to its ability to keep sources anonymous. A similar trend will prevail in the success of PSS.

While ensuring privacy, the integrity of data also needs to be maintained to provide reliable responses to queries. For example, if the fuel price at a suggested fuel station is not the cheapest, due to a loss of data quality, dissatisfied users will not participate in future. That is why it is indispensable to maintain data integrity at the intended receiver's end. Ultimately, the best case scenario is achieving anonymity at the adversary's end and data integrity at the service provider's end. Because ultimate service can only be provided when the service provider can interpret reported data at a much higher accuracy. Data integrity is undoubtedly orthogonal to security/privacy. A further point is no one can maintain any sort of privacy from an entity which is supposed to provide full data integrity. Hence, finding an acceptable privacy-integrity trade-off is crucial to ensuring voluntary critical mass participation. This privacy-data integrity trade-off model should be such that either the privacy requirements will dictate achievable data integrity level or vice versa.

Privacy concerns may be violated by two types of attacks. First, a malicious node of the participatory sensing network may abuse its ability in decrypting data to compromise the payload being transmitted. Secondly, a third party adversary not having the ability to decrypt data payloads may eavesdrop the wirelessly transmitted data and track the traffic

flow information hop-by-hop. For instance, when a malicious third party knows that a person is in a mental hospital, she is able to infer that the victim has a mental problem which is a severe breach in the victim's privacy [42]. Moreover, in the ubiquitous computing environment, it is easier to intercept the message than wired networks. The attacker can easily acquire location information without the consent of a user from intercepting messages and then inferring the context of the victim by collecting and analysing the victim's location information.

Some prior information about the target victim strengthens an adversary. It is natural that an adversary has close access to a victim in the real world and thus knows a victim's user id. Then, overhearing partial information beforehand, an adversary may decide her strategy, i.e., where to position. For instance, if someone overheard a conversation of her colleague that he will visit a doctor next Friday afternoon there may be privacy issues. Moreover, if she receives some information from an eavesdropped message that her boss was near a particular hospital at that particular time, she may deduce that her boss suffers from a particular disease which is being treated in that hospital.

2.1.2 Location Privacy in PSS

The uniqueness of PSS lies in its data communication infrastructure which is constituted by the deliberate participation of a community. The potential lack of privacy of the participants in such system however makes it harder to ensure their voluntary contribution. On the one hand, preserving the privacy of the individuals who contribute data introduces a key challenge in this area. On the other hand, data integrity is critical to making the service trustworthy and user-friendly. Different interesting approaches have been proposed so far which protect the privacy that will in turn encourage the participation of the owners of data sources.

The literature suggests that there are two main types of privacy concerns: data-oriented and context-oriented. Data-oriented concerns concentrate on the privacy of data collected from or a query posted to a participatory sensing system. Context-oriented concerns instead focus on contextual information such as the location of the participant and the timing of the traffic flow in a participatory sensing network. Another privacy issue of associating the user and the data evolves the concept of ownership privacy, as the mobile user may not want to release the ownership information of a controversial contribution. On the one hand, ownership privacy preserving schemes indirectly inherit the virtue of maintaining a user's

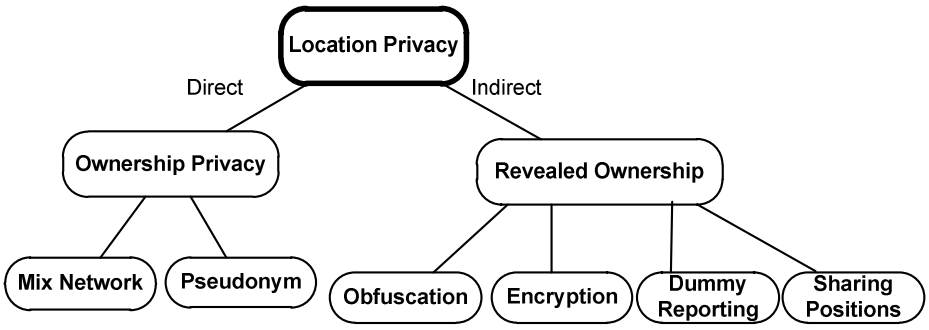


Figure 2.3: Taxonomy of location privacy in participatory sensing.

location privacy. On the other hand, revealed-ownership concerns focus on revealing ownership with the reported data at the service provider end to ensure reward eligibility and reputation score. This type of location privacy requirement is dealt within the common approach of confusing the attacker. Most of the existing location privacy preserving techniques are based on the concept of *k*-anonymization which can be achieved with or without the help of a trusted third party. In cryptography, a trusted third party is defined as an entity which facilitates interactions between two parties who both trust the third party while the third party reviews all critical transaction communications between those parties. Moreover, combining a number of the privacy protection concepts also serves as a preferred approach to preserving the location privacy for some researchers.

Balancing all the privacy requirements is a cumbersome job. Owing to a various applications, the priority settings for privacy of the participants may be different. In most cases, location privacy is likely to be the main concern of participants. While in some others, the participants may prioritize their ownership privacy or contributed data privacy.

Location privacy preservation in the context of participatory sensing has some similarity, with the same problem in *ubiquitous* computing and *ad hoc* computing areas. Existing techniques to deal with it are mostly based on *k*-anonymity, pseudonym, or obfuscation concepts. The concept of *k*-anonymity states that a data or query collected by an application is *k*-anonymous if it is indistinguishable, with respect to some chosen attributes, among $k - 1$ other data or query received by the same application. It is mostly used in Spatial cloaking schemes, although it suffers from a lack of data integrity.

Section 2.1.2.1 presents a brief review of the literature that uses the concept of mix network (throughout the thesis another term friend network is used interchangeably). In

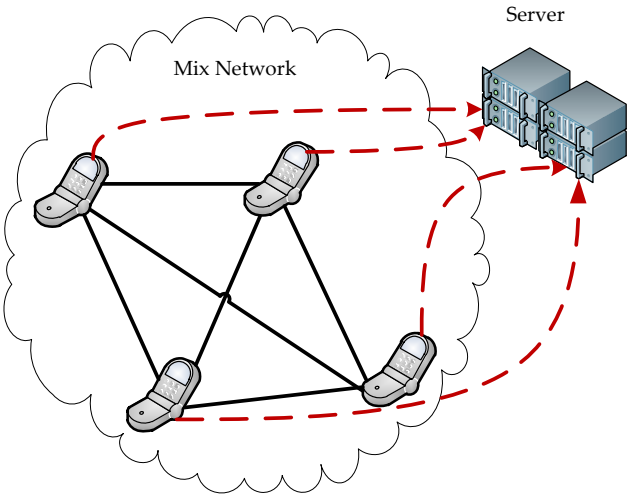


Figure 2.4: Basic working principal of a Mix Network.

Section 2.1.2.2, pseudonym-based approaches are presented followed by obfuscation-based approaches in Section 2.1.2.3. Section 2.1.2.4 outlines some encryption-based privacy preservation schemes. Section 2.1.2.5 examines dummy-based anonymization techniques. Finally, in Section 2.1.2.6, privacy preserving schemes are critically discussed where anonymization is achieved through sharing location information.

2.1.2.1 Mix Network

The basic idea of mix network, as illustrated in Figure 2.4, is to route the data through the network for a specific hop count before sending it to the ApS. This concept is used to address ownership privacy [43]-[55]. It is also preferred for applications where data with more geographical information is needed and, hence, rejects other methods. This very concept is conceived from the typical idea of layered routing, commonly known as onion routing. To provide low-latency anonymity, a second-generation onion router, Tor was designed in [44] for the TCP-based applications. With around 500,000 users [45] and approximately 3000 voluntarily-run nodes [46], it is the most widely organised anonymous data communication system. Many researchers were attracted to work on some of the major performance degrading factors of this traditional scheme which included:

- service-delays from data relaying [47] or volunteer nodes with slow connections [48]
- trustworthiness of the nodes [49]-[51]
- no flexible tuning option between performance and anonymity [49]-[52]

- no intelligent path selection [53]
- low performance video streaming [54]
- designing proper incentive for volunteer nodes [48].

Using the mix network concept a Privacy Assurance system for Mobile Sensing Networks (PA-MSN) was designed in [30] to protect ownership privacy and in turn, protect location privacy. A mix network based Hot-Potato-Privacy-Protection (HP^3) algorithm was designed in which user sent the data to one of the network friends (mix node) and that friend would choose another friend to deliver the data to the next hop. The last user sent the data to the server when the pre-defined hopping threshold was reached. The possible adversary may be a malicious server or a compromised peer mobile user. To address malicious users, the data was encrypted using the server's public key and the communications between friends were secured by some pre-negotiated shared secret between each pair of them. However, this approach can be risky for friends as they are unaware of the content and may be transferring data of anti-social activities. Due to the limited number of hops everyone became a suspect for a malicious server and it could not make an attack on the user privacy with a probability greater than $1/n$ even with the full knowledge of the network. Here n is the number of total registered users. To address the problem of interception by compromised nodes, they extended their algorithm and introduced image splitting and redundancy, where each piece took individual paths to reach the server. The strength of their work lies in addressing two privacy issues- location and ownership at the same time, which considers two attack models- malicious server and corrupted user. However they did not evaluate the cost of redundancy or image splitting. Moreover they did not report on the optimum number of friends, which is important because vulnerability may arise from having very few friends. They assumed that the data collection server often logged user identities along with the data they reported which may lead to disclosing sensitive information by compromising user privacy.

Wang and Ku [43] have suggested that most of the existing works on privacy preservation by anonymity have not been designed specifically for mobile environments and, thus, ignore issues raised by resource constraints. Even in [30], the data itself was routed through multiple users, volume which causes large consumption of network bandwidth. While the proposed method, One-way utilized peer-to-peer networks, aims to facilitate anonymous data transmission to protect user identity, the actual data payload is

not sent through peers. The procedure [43] is interesting enough to explore as an alternate option for [30]. In describing the method, the parameters and equalities have been borrowed from [43]. Here, it is only the connection request, c that was sent through the peers. It consisted of the connection identifier, I_s along with the receiver's routing address A_r and a particular hop count h, i.e.,

$$c = \{I_s, A_r, h\} \tag{2.1}$$

where, $h \geq 10$ and I_s was randomly generated and persisted only during transmission of the sensed data. After travelling through h hops, when c arrived at the server, an acceptance message a was sent back consisting of a connection token I_r, with the connection identifier, I_s following the reverse path of the peers that took part in forwarding c. So,

$$a = \{I_r, I_s\}. \tag{2.2}$$

Upon receiving this acceptance message by the data originator, the actual payload was sent directly to the server that contained on top of the actual data D, the receiver's routing address A_r, connection identifier token I_r, and sequence number s_i to identify individual packet of a transmission. Hence, the payload packet p was formulated as,

$$p = \{A_r, I_r, s_i, D\}. \tag{2.3}$$

Thus, through the whole process, the One-way protocol successfully avoided sender identity information and yet simultaneously confirmed a secure data transmission. Moreover, in this anonymous data transmission scheme the actual data was sent directly, avoiding the multiple time data replication through the peers and, thus, saving unnecessary bandwidth consumption.

Hsiao *et al.* [55] has pointed out a major drawback of a mix network is that a communication design with stronger anonymity has to bear the cost of a higher latency. From a service point of view, users want to have an intermediate privacy level without sacrificing latencies. Keeping that in mind, a practical network-based solution, Lightweight Anonymity and Privacy (LAP), was proposed in [55]. Here, each packet carried its own forwarding state, whereas all the relay nodes in the circuit were predetermined in Tor. It used encryption schemes for each Autonomous Domains (AD) to use a secret key to encrypt

and decrypt forwarding information in packet headers. Hence, an AD's forwarding information was kept hidden from all the other entities, while an LAP packet remained the same at each hop. It also designed provision for an end-host to trade anonymity for improved performance, with different privacy levels available. Here, they assumed a weaker attacker model, considering the attacker can compromise any AD except the first hop AD, where the victim end-host resides. Because of this relaxed threat model, LAP was only appropriate for users with trustworthy local ISPs demanding protection from being tracked by Websites and ISPs that are further away.

2.1.2.2 Pseudonyms

Another option for preserving ownership privacy is Pseudonym-based communication by the user [56]-[62]. For scenarios where locations are public and visited by many people this pseudonym-based method is more appropriate as anonymization approaches greatly decrease the data quality in these situations. In contrast, space cloaking technique is not a preferable solution here as data with more geographical information is required.

Beresford and Stajano [56] have concentrated on the class of location-aware applications that accept pseudonyms, and thus ensure anonymization of location information. A long-term pseudonym for each user cannot provide much privacy. Hence, the framework was based on frequently changing pseudonyms and, thus, users avoided being identified by the locations they visit. This framework was further developed by introducing the concept of mix zones as a connected spatial region of maximum size in which none of the users registered any application call-back. In other words, a mix zone concept was applied in the scene whenever two users occupied the same place at the same time. Thus, it provided unlinkability between users coming in and going out of the zone. However, a powerful adversary may use historical data to de-anonymize pseudonyms more accurately. Then again, a smaller mix zone may not solve the anonymization problem in practical situations. The study also demonstrated that even with a relatively large mix zone, location privacy can be low due to a high temporal and spatial resolution of the location data generated by their applied system.

The mix zone concept was used in the *MobiMix* approach as proposed by [60]. Zhong and Hengartner [61] proposed another option for preserving location privacy when requesting a location-based service using pseudonym-based communication by the user. The study introduced the idea of multiple servers to trade-off between centralized and

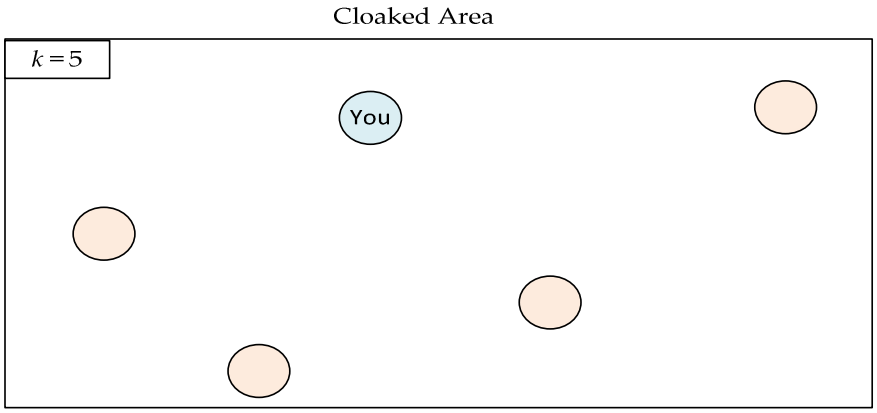


Figure 2.5: Basic idea of Spatial Cloaking to achieve *k*-anonymization.

distributed systems. The multiple servers were owned by different organizations that deployed location brokers to keep track of the current location of users. Gao *et al.* [62] improved this theoretical mix zones model using the time factor from the perspective of graph theory. The study identified that most of the researchers focused on the participator's location privacy whereas their trajectory privacy remained not much taken care of. They proposed a Trajectory Privacy-preserving Framework (TrPF), with the main strength of TrPF that it can afford trajectory privacy with lower information loss and costs than that carried by other proposals.

As explored in [58], a combined analysis of reported data with corresponding reporting patterns may help to detect the users' residences among other noteworthy places, e.g., workplaces and favourite entertainment centres, from their location traces. Hence, further measures to preserve location privacy must be added on top of pseudonyms.

2.1.2.3 Spatial Cloaking or Obfuscation

The concept of obfuscation recommends that location privacy can be preserved by intentionally reducing the precision of the location information as used in the communication [63]-[81]. In some literature, the spatial obfuscation is referred to as spatial cloaking, as illustrated in Figure 2.5. The common essence is to hide sensitive location information, making communication wilfully ambiguous and harder to interpret. Most such techniques apply the concept of *k*-anonymity to make the location information confusing. The meaning of *k*-anonymity is to make an entity indistinguishable among $k - 1$ other similar entities as introduced in the area of location privacy by Gruteser and Grunwald [64]. It is a very widely-used privacy preserving approach not limited to location privacy only.

The model for *CliqueCloak* [63] accommodated different *k*-anonymity requirements for each user, but actually compromised real time operation as it waited until *k* different queries had been sent from a particular region. Path Confusion [65] incorporated a delay in the anonymization and just like *CliqueCloak* it compromised real time operations. In *CacheCloak* [66], mobility prediction was made to enable prospective path confusion while using potentially not-trusted ApS. Mobility predictions were entirely based on previously observed user-movements and therefore prevented predictions from absurdities such as passing through impassable structures or going the wrong way on one-way streets.

Obfuscation was first introduced in [67] as a new technique to safeguard location privacy. It aimed to achieve location privacy by providing imperfect spatial information. To degrade the quality of information about a person's location, instead of providing a single position, a set of locations was sent to the location based service provider. However, too much imperfection of information eventually degraded the quality of location-based service. This limitation has been addressed in this study by introducing the idea of automated negotiation to achieve desired balance between the level of privacy and the quality of service. Nevertheless, selecting an appropriate obfuscation set was a difficult task. This is particularly true as in a cloaked region; the obfuscation set was a discrete one which incurred high communication costs to send to the server.

This idea of obfuscation was enhanced in [68] by introducing a composition of various obfuscation techniques. The study established the concept of relevance as a general functional metric for location accuracy that qualified a location with respect to either accuracy or privacy requirements. As a cloaked region, the obfuscation set was assumed to be planar and circular. Henceforth, the three possible varieties of obfuscating a circular set were obfuscation by enlarging the radius, by shifting the centre, and by reducing the radius. To satisfy users privacy preferences, any one among the three, or a composition of any two techniques, could be applied.

Anonymizing the location of a query source and processing the transformed spatial queries was focused in [70]. It discarded *CliqueCloak* for compromising real time operations. Two approaches of *k*-anonymization by spatial cloaking have been proposed—Nearest Neighbour Cloak (NNC) and Hilbert Cloak (HC). NNC addressed the centre-of-ASR attack, but could compromise spatial anonymity in the presence of outliers. A theorem was then established for a spatial cloaking algorithm to guarantee spatial *k*-anonymity if every tile

satisfied the reciprocity property. HC satisfied the reciprocity property which stated that a tile should contain the user and at least additional *k*-1 user and every user in that tile also will generate the same tile for the given *k*. It was achieved by utilizing Hilbert space-filling curve to transfer multi-dimensional data onto one-dimensional space. The adversary model stated that ApS might be compromised or it might reside in between AS and ApS. However, the weakness of this work was that the AS was assumed to be a trusted server which is not very practical and also may cause a single point of failure.

Spatial cloaking was improved in [71] by partitioning the domain into a number of safe and as small as possible subdomains ensuring each subdomain contained *k* users and that each node took the subdomain it resided in as its cloaking area. The authors provided an analytical model of communication overhead to find this cloaking area. However, their approach was not compared in terms of cost or size of the cloaking area with similar works in literature. This subdomain concept was somehow similar to Tessellation [72] which was basically *k*-anonymity by generalization. It involved partitioning the geographic area into a collection of cells and merging neighbouring cells to form tiles that users could use to mask their true locations. This concept of tessellation was introduced in [72], a novel blurring mechanism to propose a framework for nodes to receive tasks anonymously. It involved clustering to protect users' privacy against the system while reporting context, and *k*-anonymous report aggregation to improve the users' privacy against applications receiving the context. Here, *k*-anonymity required that at least *k* reports were combined together before being revealed.

Almost at the same time, the concept of *l*-diversity was introduced in [73] to address the limitations of *k*-anonymity based techniques while handling some specialized attacks. They showed two types of attacks with which a *k*-anonymized dataset might encounter severe privacy problems. First, they showed that an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. Second, attackers often have background knowledge, and then *k*-anonymity does not guarantee privacy against attackers using background knowledge. They gave a detailed analysis of these two attacks and proposed a novel and powerful privacy preservation scheme called *l*-diversity. To ensure *l*-diversity, every group of tuples that shared the same non-sensitive information should be associated with at least roughly equally distributed sensitive values.

Huang *et al.* [76] enhanced the concept of Tessellation for preserving spatial and temporal privacy in the context of participatory sensing. To protect identity disclosure of the transmitting user information, they adopted *k*-anonymization and to guard against attribute disclosure, they implemented *l*-diversity. They made an important contribution in identifying the significant problem of protecting participatory user's location privacy, which is supposed to be important from user's viewpoint although her association with the reported POI needed to be revealed to ensure data integrity. However, this work did not consider the potential damage of data integrity by their proposed solution where users from different points of interest report the centre of a tile consisting of k points of interest or alternatively their mean location as location of all those different points. The receiving server associated this reported point with the nearest point of interest and, thus, suffered from false association of $k-1$ other points within the tile. Data integrity is somehow orthogonal to security/privacy. Hence, finding an acceptable trade-off is a challenging task. Moreover, for anonymization purposes they depended on a third party entity, and therefore, the AS might suffer from the limitation of a single point of failure or being compromised. The limitation of Huang's naive algorithm we have identified was also addressed in [70] as they described it as a centre-of-ASR attack.

Similar to our research focus, Rodhe *et al.* [79] investigated the impact of privacy techniques such as *k*-anonymization that introduce uncertainty in data and on the quality of information at the receiving end. Using two strategies to reconstruct the data distribution, they found that the cloak area resulted from applying *k*-anonymity which influences data quality more than the size of the anonymity set (*k*). However, the anonymization techniques they were using assumed a simple model of data flow to the destination server where the identification of an observer was not available with the observation. The implication of this model was that no user specific reward-scheme or reputation mechanism was applicable and, thus, created a barrier in making PSS popular.

Another dimension to spatial cloaking was given in [74] by introducing *l*-diversity along with *k*-anonymity and incorporating temporal cloaking functionality into the location perturbation process. At the same time, the study maintained the user's preference for privacy to accommodate dynamism. One shortcoming of this approach was that there was a communication overhead that incurred for every mobile user participating in the process. However, this factor could be balanced out considering that the anonymization success rate

and desired service was achieved. Then again the dependency on a central trusted server made it vulnerable in real-life scenarios.

A new framework named *Casper* was proposed in [69] that consisted of two main components, the location anonymizer and the privacy-aware query processor. It provided location privacy for a query source accommodating user specific anonymity preference. Here, the location anonymizer blurred the location information by spatial cloaking. Then the privacy-aware query processor, which was embedded in the ApS, gave a candidate answer list that was inclusive and minimal. It could be applied to a large number of mobile users. However, it suffered from the dependency on a trusted third party.

The problem of query privacy was addressed in *SpaceTwist* [75] by applying obfuscation to generate an anchor and retrieving information on k nearest points of interest from the location based service provider. This scheme needed neither any trusted third party nor any communication between other users to form groups. It maintained a good balance between the privacy of the user and the success rate of finding the closest point of interest as a reply of the query. Nevertheless, it did not deal with the dynamicity of any user's privacy preference and only concentrated on obtaining k nearest neighbour replies successfully.

To address the bottleneck of a centralized trusted third party Chow *et al.* [77] introduced a distributed system architecture and proposed the first peer-to-peer (P2P) spatial cloaking algorithm to protect location privacy of mobile users. In this algorithm the user achieved ***k-***anonymity by collaborating as a group with other $k-1$ nearby peers. Thus, it approached the mobility of the user. One of the peers from the group then acted as the agent and forwarded the query on behalf of the originator. As the query was based on the cloaked spatial region, the location based service provider provided the agent with a list of candidate answers which was readily forwarded to the originator. The query originator then acquired the actual answer filtering out the other false candidates. Here k-anonymity was a user specified privacy requirement which was a key aspect of this algorithm. Nonetheless, this approach lacked an ability to highlight how to deal with potentially compromised peers

In the field of ad-hoc network, Hashem and Kulik [78] contributed the same interesting idea of using one of the $k-1$ other mobile nodes to act as the query requestor to protect the identity of the query initiator and, thus, maintained the location privacy of the user. At the same time, the study addressed the short-comings of previous work by coping with no trust among peers. It enhanced the previous idea of [77] by combining anonymity and

obfuscation, thus addressing the potential security threat of trusting a large number of group members. At first, obfuscation was employed by each mobile user to hide her actual location in a Locally Cloaked Area (LCA), both from the service provider and her peers. Then anonymity was achieved by combining its LCA with the LCAs of $k-1$ other peers. Finally, the *k*-anonymized Globally Cloaked Area (GCA) was sent as the location information, while requesting a location based service from the service provider. Thus, this approach was totally free from trusting any of the involved parties, either the peers or the service provider. However, communication among the neighbours and the service provider might increase the overall communication costs of this technique. Group queries offer a new dimension of privacy challenges as here the location of all group members are vital to discovering their nearest neighbour while at the same time any group member can be compromised. The concept of private filter was developed in [80] to determine the actual group nearest neighbour without revealing the user's location to any involved party. This was set against the queries from a spread out group of users providing their locations as regions instead of exact points to the service provider. However, this approach did not work for group of size 2. This is an interesting problem with similar issues for participatory sensing applications.

Vu *et. al.* [81] presented a spatial cloaking based *k*-anonymous location privacy technique for participatory sensing applications. They emphasized the quality of spatial cloaks indicating that the cloaks should be close to the user's location and small in size so that search algorithms can be executed efficiently. From this motivation, they devised a mechanism based on Locality-Sensitive Hashing (LSH) to partition user locations into groups each containing at least *k* users. They then proposed another algorithm to answer queries for any point in the spatial cloaks of an arbitrary polygonal shape.

2.1.2.4 Encryption

Encryption may seem to be the most obvious solution when the question of security and privacy arises. It is already an established measure to provide data privacy. However, the context and severity of breaching privacy also needs to be considered when the expenses involved in encryption are involved. Moreover, data encryption cannot do much to preserve location privacy. In most of the cases [82]-[90] encryption was employed only to assist the main location privacy preserving method. Takabi *et al.* [82] introduced a cryptographic scheme to adopt the distributed collaborative approach to achieve location privacy based on

k-anonymity. It required neither a trusted third party nor the users to trust each other. Cryptography was used to learn the presence of minimum k users in the query area including the query originator. Thus, it replaced the need of the location broker from Zhong's [61] solution. Though both the methods exhibit efficient implementation, a comparative study of their performance could have made it more worthwhile.

In [83], a privacy protecting layer using cryptographic tools was proposed, where no entity excluding the network operator can learn the current location of a mobile node. Most of the encryption based proposals focused to target on avoiding the bottleneck of depending on a trusted third party. The encryption based approaches follow the cryptographic techniques and terminologies.

The concept of multi-secret sharing [84] states that some arbitrarily related secrets can be shared among a set of participants who are not trusted individually. A number of location privacy preserving schemes [85]-[87] adopted this concept to propose a novel position sharing approach, while taking service from location-based applications. In the very recent [87], the multiple shares of position information were generated and distributed among a number of location servers. Then the location service providers, i.e., the clients, combined the specific shares obtained from specific location servers they got access permission from the user. The main target of this cryptographic approach was to develop *secure* share generation and combination algorithms to defeat malicious clients or location servers by the level of precision. In describing the method, the parameters and equalities have been borrowed from [87]. According to their share generation algorithm, the MN splits up π into a *master share* m_π, and set $S_\pi = \{r_{\pi,1}, .., r_{\pi,b}\}$ of b refinement shares by calculating

$$generate\ (\pi, l_{max}, b) = (m_\pi, S_\pi), \tag{2.4}$$

where l_{max} denotes the number of different precision levels. Then the clients received permissions to access a user-defined set $S'_\pi \subseteq S_\pi$ and executed share combination on the public m_π and these refinement shares using,

$$combine\ (m_\pi, S'_\pi) = p\ (\pi, l), \tag{2.5}$$

where $p\ (\pi, l)$ defined the obtained position with precision level l. It was established that if the known precision of position π_{attack} was $precision(\pi_{attack})$, then

$$precision\big(p(\pi, l)\big) \geq precision(\pi_{attack}). \quad (2.6)$$

The scheme was explored both for symbolic and geometric location models. It assumed that the location servers were always online, and that the user can permit access to some trusted applications. To access this information, the clients took reference of some pre-shared secrets requiring the overhead of sender/receiver pre-shared secret setup.

The concept of the Private Information Retrieval (PIR) technique was used in [89] to answer queries without extracting any information from the query. Symmetric encryption techniques were used in [88] to share a secret with the friends. Here the approach was developed to inform users of the presence of their friends within their proximity without revealing the location of the user to the ApS. Cristofaro *et al.* [90] addressed privacy threats as posed by smartphone applications that provide service on the basis of users' personal information and preferences. For this, they proposed using a semi-trusted server which deals with encrypted individual data input. Both the server and participants were assumed to be acting rationally, i.e., not trying to fail the system. However, the cryptographic operations were not optimized. Moreover, the performance analysis was done to compare different versions of this very scheme and lacked comparison with any state-of-the-art practices.

2.1.2.5 Using Dummies

The idea of using dummies states that the location information will be privacy-preserved if it comes with multiple false location information or false data traffic i.e., includes dummies and confuses the adversary thereby. In one of the initial approaches, diffusion method was employed in [42] that scattered the user's location information to confuse the attacker. In addition, the base station or the access point transmitted dummy messages that looked like real traffic but had no actual meaning. However, the diffusion was not applied to multiple packets, and the dummy user was not made more like a real user.

Generating the dummies can be done with or without the help of a trusted third party. In [91], the user herself generated the dummies without the help of any such external entity. The common challenge in working with this concept is creating realistic dummies to make the actual data ambiguous. A fully decentralized and autonomous *k*-anonymity based client-side system, *SybilQuery* for preserving privacy of location-based queries was presented in [92]. The basic framework demonstrated that for each query from the client it would

generate $k - 1$ dummy queries and, thereby, would ensure k-anonymity. It addressed the limitation of most of the spatial cloaking approaches which depended on a trusted third party anonymizer. Peer-to-peer techniques (e. g. Friend Network), on the other hand, relied on the participation of k peers and, thus, restricted the autonomy of such systems. The approach proposed here required no change on the server side and only required minor modifications to the querying clients. In the implementation, as input, it took a path to be followed by a vehicle along which a vehicle might issue several queries to the ApS. It would then output $k - 1$ dummy paths that statistically would resemble the input path. To make it more appropriate in practical situations, the basic design had accommodation for some extensions like randomizing path selection, handling active adversaries, endpoint caching, providing path continuity, and adding GPS sensor noise. Thus, the efficiently generated queries were indistinguishable from the real ones. The computational costs of this system mainly consisted of handling different databases involved in the different steps.

2.1.2.6 Sharing Location Information

The concept of location information sharing states that a number of user nodes will share the location information and produce the combined data to confuse the attacker. This concept was used in [85] to securely manage private location information in an untrusted system. This approach was further extended in [93] by including map knowledge into account to prevent adversaries attempting to increase the precision of location information. Some of the methods used geometric transformation to generate the shares whereas some [87] used the multi-secret sharing concept [94] to do the same. Both approaches support symbolic location information.

Boutsis and Kalogeraki [95] focused on developing a PSS for android smartphones and proposed an efficient low cost and distributed approach for users to disclose their trajectory info without compromising privacy. The proposed approach assumed user data would be stored locally on the individual smartphone devices, i.e., no dependency on any centralized database. Apart from expenses involved and a single point of failure, their argument for using a distributed approach is that the users typically query data that is dependent on the location of the user. To maintain privacy by making all local trajectories equally-probable to be sensitive data, the data exchange approach distributed user data trajectories among multiple user databases, based on local entropy. They considered different types of attacks

including those arising from the use of Android OS itself, as well as user identification attacks, sensitive location tracking and sequential tracking attacks.

In [96], [97] location privacy was preserved during the sensor reading collection by applying a collaborative path hiding mechanism in a decentralized way. Instead of directly reporting the location to the application server, this collaborative path hiding concept was applied to isolate the spatiotemporal context (i.e., time and location), at which the sensor readings were taken, from the identity of the users by physically exchanging location information between users in an opportunistic approach. Three different exchange strategies were proposed in [97] to explore the impact of different levels of symmetry of each exchange that cluttered the location information between users. Each user uploaded every hour a combination of the samples received from other users and the remainder of their own collected samples to the application server. Finally, by analysing the reported location information the application server built summary maps of the occurrence being observed (e.g., noise pollution map) to provide service to the public. Thus, this approach mainly focused on maintaining location privacy while contributing to a participatory sensing application that works on the structured aggregation of collected data without real-time constraints for data delivery. Hence, it could safely overlook issues like real-time data fluctuation, pinpointing the exact location of a desired sensory data etc.

2.1.3 Data Privacy in PSS

There is a variant of this problem named data privacy, which is the privacy of individual data or ownership of sensitive data. Talking about privacy in data communication, the first thing to come to our mind is encryption. In cryptography, encryption is the process of converting information using an algorithm to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. Encryption however, is not viable in participatory sensing systems as in these cases the sensed dataset is usually small and predictable to defeat public key cryptography. Moreover, law enforcement agencies require public network dataflow remain unencrypted to mitigate national security threats.

Data oriented privacy is another area of research interest which is relevant for many data collection types. Reddy *et al.* [98] have suggested that PSS can support several types of data collection such as that initiated by some researchers or from the ordinary citizens themselves. The main challenge was to allow individuals to share data for computing community statistics with a privacy assurance. This is because they may be interested in the

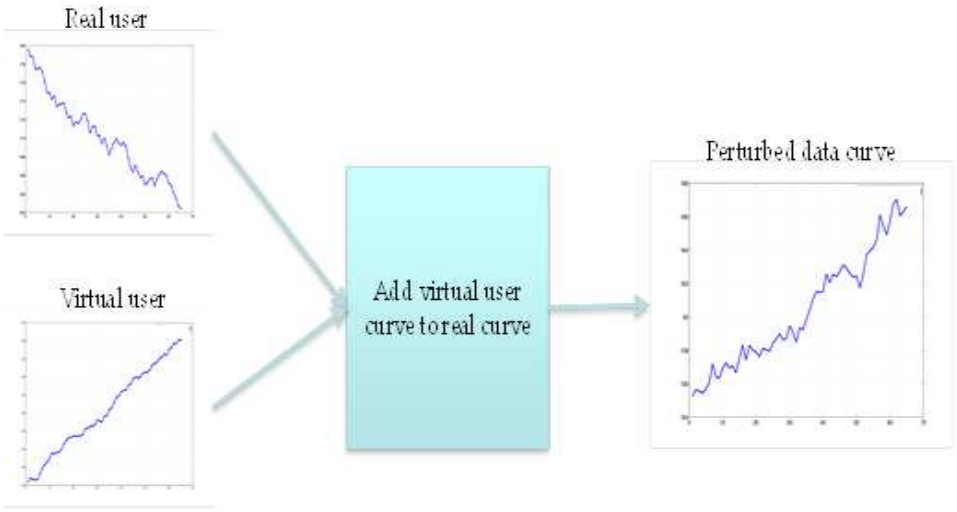


Figure 2.6: Basic idea of Data Perturbation.

statistics, but do not trust sharing their private data with the third party. For example, in recording her weight periodically a weight watcher is expected to be interested in knowing the efficacy of a diet chart, irrespective of her preference to hide her true weight and/or average weight and/or trend of weight i.e., loss or gain. The user may want to find the average weight loss trend as well as the distribution of weight loss as a function of time on the diet from which individual's weight and/or weight trend should not be extractable.

A very common approach to preserve data privacy is to apply the concept of data perturbation to add artificial noise to the data (see Figure 2.6). To prevent adversaries from reconstructing the individual original data, independent random noise was demonstrated to be insufficient in [99]. In *PoolView* [100], mathematical foundations and architectural components for providing privacy guarantees on stream data in grassroots participatory sensing applications was developed. It relied on data perturbation at the data source to allow users to ensure the privacy of their individual data as they used tools that perturb such data prior to sharing for aggregation purposes. It then used community-wide reconstruction techniques to compute the aggregate information of the interest. Thus, user privacy was preserved against traditional attacks, like filtering and specialized attacks such as MMSE, and at the same time community information both the average and the distribution were successfully recovered. This approach is best suited for a closed community with a known empirical data distribution.

This privacy preservation model was further enhanced in [101] to ensure the correct reconstruction of community statistics in the case of correlated multidimensional

time-series data. It also proposed a perturbation-based approach of addressing data privacy in participatory sensing. The system was applied to construct accurate traffic speed maps in a small campus town from the shared GPS data of participating vehicles, where the individual vehicle were allowed to "lie" about their actual location and speed at all times.

Privacy preserving data aggregation has been gaining popularity mainly in the field of sensor networks. This field differs from PSS in a way that sensors are deployed with a single authority and their static topology conflicts with the dynamic property of mobile users that constitutes the data infrastructure of a participatory sensing system. Being inspired by this portion of work in sensor networks, *PriSense* was proposed in [102] to support user privacy in data aggregation. This technique has mainly been based on data-slicing and mixing. The initial idea was very close to Privacy-preserving Data Aggregation (PDA) concept of [103]. However, the main difference lies in their application scenario and how the cover nodes were selected.

As introduced in [104], *k*-anonymity model is equally popular in protecting data privacy. Traditional *k*-anonymity based privacy preserving approaches [105], [106] were extended in [107], used for pattern mining in [108], [109], and for real-time social network data sharing in [110].

Choi *et al.* [111] addressed the privacy threats of sharing highly personal information via participatory sensing applications which provide data to medical behavioural studies or personal health-care schemes. The proposed architecture, *SensorSafe*, used an access control mechanism with numerous privacy preference options to allow users' control over the behavioural information they wanted to share. The concept of broker was used here to support data management from multiple contributors. However, neither any backup plan in case of a single point of failure nor any sort of trust assumption of this broker was mentioned. Moreover, no adversary model or performance analysis in the context of communication overhead or operation efficiency was made which could have added more value to this study.

Preserving privacy is an even more challenging job when multidimensional data is fed to participatory sensing applications. Keeping that in mind, the concept of negative surveys was used in [112] to facilitate the complemented sensory data as an input to their proposed algorithm to work with and get aggregated result of the actual ones without revealing actual individual data. This negative survey concept was preferred over the computationally

expensive encryption and key management schemes to serve the very purpose of privacy and security in sensitive data handling. Furthermore, they acknowledge that encryption is not applicable in cases where even the intended recipient is not trusted with the ownership of sensitive data. The dimension adjustment concept was introduced to recover the limitation of a huge increase in the required number of participants to maintain a given level of utility. Besides they explored the somewhat orthogonal relationship between privacy and accuracy.

Cristofaro and Pietro recently proposed some techniques to ensure query and data privacy in urban sensing systems [113], [114]. They explored different adversarial models depending on whether one has control over a fraction of sensors before or after the data being sensed and also considering whether the adversary is randomly distributed or local to a specific region of the network. For each of these settings, they presented a probabilistic distributed technique that trades off an achieved privacy level with a potential communication overhead.

2.1.4 Other Related Works

Gadzheva [115] suggested that protecting privacy in an increasingly transparent society will only be possible with the development of an intelligent interplay between technological design and legal regulation that reflects the great expansion of ubiquitous data processing and surveillance capabilities. Failure to address the legitimate concerns of users can have a negative impact on businesses, network operators and service providers, and seriously impact the deployment of these beneficial services.

Christin and Hollick [116] investigated the technological basis for mobile sensing applications with an analysis of different sensor modalities collected in existing applications. They wanted to highlight the respective threats to privacy. Their main focus has been to examine how the sensor readings are processed within a range of typical mobile sensing applications.

In this practical world, users are generally interested in participating if only they can benefit themselves or understand the benefits for a wider community [117]. Otherwise the users may gradually lose interest in participating actively and, thus, appears the challenge of identifying methods to motivate user participation. To encourage selfish participants, [118] proposed a novel bargain-based stimulation among the MNs in PSS. It presented a greedy

algorithm formulated with the help of game theory. However, they ignored addressing the participant's privacy concerns.

It is not unusual that due to lack of any proper incentive scheme [119] the participants may start finding no interest in remaining active in the system. In the absence of proper evaluation of the contribution, it is quite unavoidable for a system to start suffering from inadequate participation. As the system depends on people participating voluntarily to contribute data, insufficient contribution may cause degradation in the quality of service. That is why the concept of a reward system is proposed to keep user participation up to the expected level and ensure ample data aggregation thereby.

Danezis *et al.* [120] addressed the valuation of user data with a fixed price which is a very naive approach. Moreover, their work does not distinguish between the different times of day, locations, or various situations a user may be in. At the same time, a user's true valuation differs among individuals and over different types of data. Thus, they only focus on obtaining an estimate of the value that users attach to their location data being used by third parties.

Lee and Hoh [121] identified that employing a traditional reverse auction fails to keep user participation up to an expected level. In an auction based reward scheme, participants produce their true valuation as the bid prices, which include all efforts for collecting data such as battery power consumption, device resources, and privacy. However, in that case users with a higher true valuation may drop off in a traditional reverse auction. As a result, they proposed a Reverse Auction based Dynamic Price (RADP) incentive mechanism with the idea of Virtual Participation Credit (VPC). Here, the loser was given a virtual credit for participating which was eventually used to lower the bid while considering her next bid price and increasing her winning probability in a future auction round.

One of the obvious reasons behind the huge potential of PSS is that it makes use of existing technology and infrastructure. Then again this is also the main cause of the challenge of coordinating systems designed for other purposes and provoking the use of models to infer what cannot be directly measured or sensed. It faces the challenges of data control and user participation by verifying participant context, validating samples, incorporating human contributions, and providing reputation scores for participants. It also identifies the concern of establishing data integrity while allowing participants to regulate

their own privacy and participation. To address this challenge *MobiSense* [98] proposed using an end-to-end data pipeline from collection to analysis.

The problem of verifying data received from user devices in participatory sensing was studied in [122] and [123]. The focus of this work was to ensure that the data contributed by a mobile phone indeed corresponded to the actual data reported by the device sensors. They assumed a threat model in which a malicious user or program tampered with software running on phones and corrupting the sensor data. Their solutions relied on an auxiliary Trusted Platform Module (TPM), which guaranteed the integrity of sensing devices. However, TPM-enabled mobile phones are yet to be mass produced and, as such, their solutions are not readily deployable. Moreover, the TPM is unable to detect malicious behaviour where the user may physically create interference that affects the sensor readings.

For participating devices of PSS, Huang *et al.* [124] introduced reputation scores as a reflection of the trustworthiness of the contributed data. As in the application scenario, this study considered a noise monitoring system which generated a collective noise map by aggregating measurements collected from the mobile devices of volunteers. In this specific application scenario, a continuous deviation from the group consensus may indicate an outlier. However, in scenarios like *PetrolWatch* [2] this deviation may also indicate a change in fuel price, a trend which is quite natural over the course of time.

2.2 Privacy-Preserving and Verifiable Voting

In essence, a verifiable election starts with ballots having a unique serial number. After casting the vote, part of the ballot is handed over to the voter so that she can verify that her vote is indeed counted for the intended candidate at the end of the election. Unfortunately, this simple voting process is vulnerable to vote buying or coercion. Prevention is a must in any modern democratic civilization. The election must also satisfy unconditional privacy to the voters. A brief background of voting system is presented in Section 2.2.1, followed by some related works in this context in Section 2.2.2.

2.2.1 Background

To ensure impartiality and integrity, each country has to follow certain universal standards. From literature [127]-[171], we may compile a brief overview on these.

- **Verifiability:** This is the measure to ensure the integrity of an electoral process. It implies that each voter can check if her own vote is included in the counting (individual verifiability) and anyone can check that all and only authorized votes are counted without any change at any step (universal verifiability). Alternatively, these are also termed as end-to-end verifiability that is achieved by the combination of three properties:
    - A ballot correctly represents a voter's preference (*Cast-as- intended*).
    - The vote is stored as it is cast (*Recorded-as-cast*).
    - The announced result is a correct amalgamation of the set of recorded votes (*Counted-as-recorded*).
- **Privacy:** The preference of a voter can never be revealed. The system or any external entity will not be able to establish any link between a voter and the vote she casts.
- **Coercion-resistance:** The voter cannot be forced to vote for a particular candidate or not to vote a particular candidate. Even if the receipt to verify is handed over to the coercer, or the vote processing authority colludes with the coercer, coercion is possible.
- **Vote trading prevention:** A voter must not be able to prove her vote in favour of a particular candidate and claim advantage thereof.
- **Authenticity:** Any voter or group of voters must not be able to prove a claim that an election result is manipulated when it is actually not. When verifiability is ensured by receipts, preventing a fake receipt is the most important task in this regard.
- **Integrity:** The final vote counting must match the number of voters and their preferences.

To conduct a verifiable and privacy preserving election, the following system entities and authorities need to function properly.

- **Voting Booth:** The voter should be able to perform some actions privately such as viewing candidate information, casting their vote, and obtaining receipt. At the same time, it has to be ensured that she cannot transmit any information (for

example, a long receipt number that is not possible to remember) electronically that can be used for coercion/trading.

- **Electronic Machines:** Voting machines are designed in a number of ways. In most cases, it is a general purpose computer consisting of sophisticated hardware and software [127]. It consists of a processor, memory, optionally input/output interfaces and uses some kind of operating system. Sometimes the random number generators needed to offer privacy of the voters are part of it and in some cases they are external.

- **Receipt Exchange Box:** After each vote is cast, the voter obtains a receipt to be used later to verify that the vote is counted properly. To establish unlinkability between a vote and the voter, a widely used mechanism is to use a receipt exchange box where the voter drops her own receipt and randomly picks up another one to verify later.

- **Public Bulletin Board:** A public bulletin board is assumed widely as a broadcast channel or as append-only storage with public read-access.

- **Auditors**: To run the post-election global verification process and and in some approaches to guarantee or verify the randomness of the challenge, reliable experts are appointed as auditors.

- **Authorities**: The central election authority opens and closes the election, facilitates the counting of votes, and announces the result. The poll workers assist the election authority during the election procedures, for instance, by checking voter eligibility, handing out ballot papers for paper-based systems, and verifying receipts collected by a voter right after each vote is cast electronically. Their overall responsibility is to monitor the polling environment, and supervise both the vote casting and receipt exchange.

The role of these entities in our system will be discussed in Chapter 6.

2.2.2 Related Works

Due to the wide demands of verifiable and privacy-preserving electronic voting, numerous studies have been undertaken in recent times. A brief overview of the major works is presented below. Section 2.2.2.1 describes various commitment scheme based approaches. In

Section 2.2.2.2, three-ballot based schemes are discussed followed by some Prêt à Voter based approaches in Section 2.2.2.3

2.2.2.1 Commitment Scheme Based Approaches

A number of voting schemes have used a commitment scheme, a widely used cryptographic technique, to ensure the trustworthiness of the voting protocol. A commitment scheme has to maintain two properties, i.e., hiding and binding. A basic presentation of the scheme is as follows. For message M and a random number R, a function *Commitment* (M, R) produces a commitment C. It can be checked with another function *Open* (C, M, R) that accepts as true iff $Commitment\ (M, R) = C$. The scheme hides if it finds any $\{M, R\}$ given C is computationally very difficult. On the one hand, if extracting only M given C is also difficult, it ensures stronger hiding. On the other hand, binding refers to the property that it is hard to find any $\{M, M', R, R'\}$ such that $Open(Commitment(M, R), M', R')$ is accepted as true. Pedersen's scheme [127] is an example of a commitment scheme found in the literature.

Among the notable works that uses a commitment scheme in the context of trustworthy voting, Bingo Voting by Bohli *et al.* [130] is one such that uses EVM. Before the election, EVM generates n random numbers $X_j^i$for each candidate P_i that eventually creates dummy vote pairs (X_j^i, P_i) and hiding commitments that are also generated for these dummy votes. When a voter expresses here an intention to vote for a particular candidate, the random number generator (RNG) generates a fresh random number r which is assigned to the candidate of the voter's choice. For every other candidate, EVM draws one number out of the pool of dummy votes randomly. In the receipt printed by the EVM, the candidate that was voted for is assigned a new random number r and a dummy vote is shown for the other candidates. In the post-voting phase, every voter can verify that her receipt is shown on the list which ensures that it was counted for in the tally. She can also verify that the number of remaining commitments is also present. The authors established the correctness of the receipts using commitments with a special homomorphism property. The authors enhanced it in [131] by proposing a hash chain such that each single receipt guards the integrity of all receipts previously issued. Apart from using a trusted RNG, they assumed a trusted EVM, which is a hard assumption and, hence, the applicability of this technique is very limited.

Scantegrity [23] and Scantegrity II [24] enhanced the existing paper ballot system using a widely deployed optical scan ballot and commitment scheme. Voters marked a ballot with their selections and obtained a receipt to ensure verifiability. The receipt was torn off a ballot

chit, i.e., a perforated corner of the ballot that contains a serial number written in human and computer readable forms. The voter also wrote down the randomly assigned code letter listed next to the selected candidate which varied for the same candidate in different ballots. Hence, a particular code letter did not reveal which candidate that person voted for. After the election, all voted confirmation codes were posted online, where voters may check them. The confirmation codes did not allow voters to reveal how they voted, however, if incorrect codes were posted, that could be proved. All the confirmation codes were visible on the ballot in the case of Scantegrity which confused the voters about which codes appeared on the website. However, the in-person dispute resolution process adopted by them did not scale well.

In Scantegrity II [24], the problem was overcome using invisible ink to design a dispute-resolution procedure based on knowledge of a secret confirmation code. Voters marked ballots using a special ballot-marking pen, which made legible pre-printed confirmation codes. When the ink in the ballot-marking pen was used on the ballot ovals or confirmation codes printed with another special ink, both of these darkened. However, the confirmation code ink reacted more slowly than the ballot oval ink, and hence darkened several minutes after the oval. The voter may note it on the chit since the code is visible for several minutes after being marked. Since the code is indistinguishable from its background in an unmarked oval, the voter may have it only after she has made the corresponding ballot selection. Voters can check the code online using the two serial numbers printed on the chit. These serial numbers also remain indistinguishable from the background until a decoding pen is used by the authority concerned who reveals the serial numbers using a decoding pen after the ballot is cast. This prevents voters from falsely claiming about a confirmation code obtained from an uncast ballot. However, the requirement of such special inks is infeasible in many scenarios and the procedure is conceptually very complicated.

Hover [28] is a recently proposed approach for trustworthy voting that has combined some other ideas. The Scantegrity optical-scan system [23] was followed for a voting protocol and Eperio [132] was adopted to facilitate verification. Trustees generated ballots each containing a serial number, a code letter assigned randomly without replacement, and associated candidate name using a trusted printer. A voter marks an optical-scan oval appearing beside the preferred candidate. Some ballots are randomly selected for audit which cannot be used for voting. The serial number and code are published on a bulletin

board along with a shuffled candidate list using two random permutations by the trustee. The commitments to the random permutations are also posted in bulletin board. Unlike Scantegrity [23] that publishes separate commitments to each pair, Hover commits to the full specification of random permutations that allowed using a simpler commitment scheme with substantially fewer cryptographic operations. The trustees randomly generate and obliviously print the confirmation codes on each ballot using oblivious printing [133] which overcome the risks of using trusted printer. Along with ballots, a vector of encryptions of each code-candidate association for each ballot is also generated. Semantically secure public-key encryption schemes are used here for which the decryption key is distributed among multiple trustees. Finally, using a secure multiparty protocol and the invisible-ink printing techniques developed for Scantegrity II [24], Hover proposed a means by which several parties can generate a shared secret and print it in without learning the result.

The paper-ballot based voting scheme, Punchscan [135], [136] used two paper sheets attached one upon the other. The list of candidates was given on the top page and a letter in a random permutation was assigned to her. The top paper had holes through which letters from the lower paper were visible. Each pair of pages had a short id, which the voting authority used to understand the content of each page. The voter needed to mark the letter assigned to her candidate which had to be visible on both papers. One layer of the ballot was collected by the voter as a receipt. Without knowing the content of the other part, the receipt did not give any information about the vote. Bohli [130] showed that the voter has to vote in favour of the coercer with 50% probability in this scheme, unacceptable by any standard.

Split-ballot [26] was proposed as a receipt-free commitment scheme based protocol using a modification of Pedersen's scheme [127] and the trust was distributed in more than one voting authority. This is a paper based system, although the machine can be used for vote casting. A vote consists of two ballots from each of the two voting authorities. Two from those are actually used for voting and the other two was for verification. A partial copy of each ballot was collected by the voter as a receipt. A vote tally was jointly performed by the two authorities who published all of the ballots. Implementation of this scheme was similar to that of Punchscan [135], [136] as it used two stacked papers and the top one contained holes to reach the bottom paper. The voter also gave her choice in a similar manner. The drawback of this scheme is the requirement from voters to be able to perform

modular addition from a randomly selected value. The trust is distributed among two independent voting authorities and the privacy of the voters may be breached if both of these are corrupt.

Other studies [137], [138] also use commitment scheme. Basically, the EVM commits itself to some random values by printing it on a receipt without showing it to the voter. The voter casts her vote and simultaneously enters a random number that is used later to prove that the vote has been counted for the appropriate candidate. Every candidate is printed along with the corresponding user choices on a receipt. The order of operations is crucial to prevent potential fraud by the EVM. Here, the user enters dummy values for the other candidates, then the voting machine commits, after this the voter enters the random value for the real candidate. The major drawback of these approaches is the reliance on the limited memory of humans in the voting process. It exposes the possibility of coercion, especially when there are a large number of candidates, a common phenomenon in national elections. Moran and Naor proposed another scheme [139] based on statistical-hiding using the Direct Electronic Recording (DER) that plays the part of the voting authority.

2.2.2.2 Three-ballot Based schemes

The three-ballot based or VAV voting systems proposed by Rivest *et al.*[27], [140] is a paper-based system which is conceptually simple and easy to implement. In the three ballot scheme, the voter marks three ballots in a single vote. An example of a valid vote using this

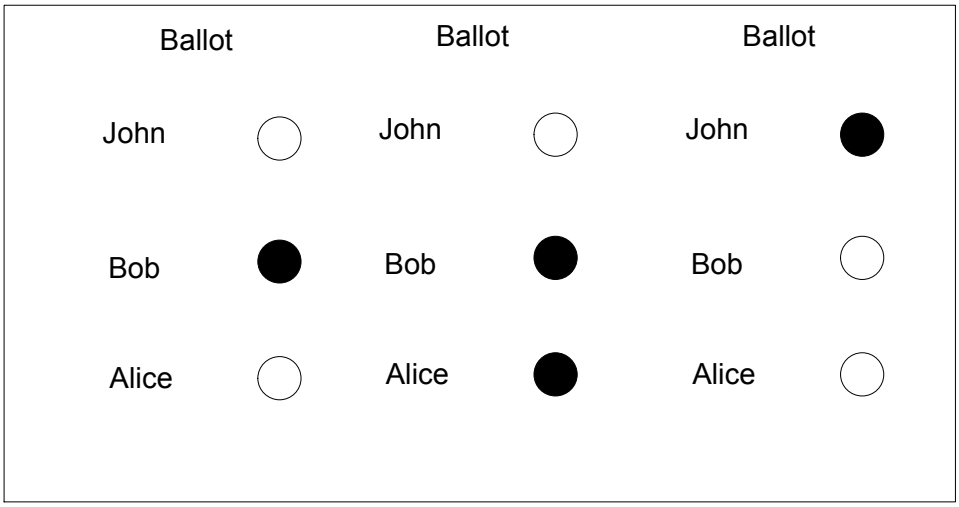


Figure 2.7: A filled out ballot in Three Ballot system with a vote for Bob. Only the row containing Bob has two filled-in circles, whereas the other rows have exactly one.

scheme is shown in Figure 2.7. The voter checks off her preferred candidate in two ballots; for all the other candidates, just one check is needed on one of the three ballots, randomly. This this system, the candidates voted for will have two marks in the three ballots set, while all the other candidates will have just one mark each. Later on, one ballot is chosen at random by the voter as a receipt. After the election, the electoral authority publishes all ballots to let voters verify whether their votes were accounted for. The remaining two ballots not taken as a receipt by each voter may be manipulated by compromised authorities. Thus, the partial information provided in receipt creates loophole for the verifiability of voting.

The authors themselves admitted its vulnerability from a three pattern attack. The coercer may force a voter to vote following a specific rare pattern. In this case, the voter will do it to avoid getting caught if the instructed actions do not appear at all in the bulletin board. This pattern encourages vote buying. Kusters *et al.* [141] have pointed out that through coercion in three ballot, the manipulator can convert the opponent's vote into the manipulator's favour as only one-third of each vote is verified. The remaining two thirds can be manipulated intelligently. This is the most dangerous problem with the three ballot based voting. Another recent attack named Clash Attack [142] pointed that the EVM may remember the particular vote pattern in the receipt a voter takes home. Next time, when another voter chooses that ballot to take as receipt, EVM puts the same receipt number in it and as desired by the manipulator replaces the rest. The authors have shown that it can work in different voting schemes.

Costa *et al.* [140] developed an electronic system adopting the concept of originally

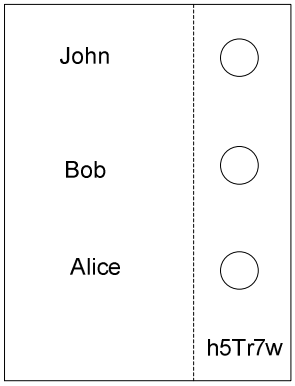


Figure 2.8: PaV ballot having two parts separated by perforation. Candidate names are printed in a permuted order. The right part has marking provision for voters and a string containing information about the permutation in encrypted form.

paper-based three-ballot system. They used an electronic ballot box, voting console, and bulletin board and also introduced a registration agent, and voting manager. However, they failed to resist the attacks discussed above that are applicable to the original three-ballot system.

2.2.2.3 Prêt à Voter Based Approaches

In the Prêt à Voter (PaV) system proposed by Ryan *et al.*[25], [145], the ballot is pre-printed with two columns. The left column presents the candidates in a shuffled order of the base ordering determined by a cyclic offset, and the right column is empty for voting. At the bottom of the right column, there is a random encrypted value, termed as onion that corresponds to the candidate order. An example of a valid ballot paper using this scheme is shown in Figure 2.8. The voter selects a ballot randomly, marks the right column, or ranks the candidates, and then tears off the left column. The right column is scanned and sent to a bulletin board and the paper is taken as a receipt. To ensure verifiability, the voter later checks if the receipt was correctly posted and thereby receives assurance that it is considered into the tallying process. However, this scheme does not provide any proof that the onion matches the candidate order in the ballot's left column or any alteration in that order would remain undetected. Furthermore, if the information shown to the voter is changed by a compromised machine (to be elaborated on in Chapter 6, as group interchange attack), there is no way to detect it.

In [145], Ryan *et al.* carried out a comprehensive threat analysis on PaV and some enhancements have been proposed by Xia *et al.* [146] that address several additional issues to handle various election models. Graaf proposed another scheme [147], combining the advantages of PaV [25] and Pushchan [135]. Due to its simplicity, it used the widely convincible ballot layout of PaV and preferred commitment primitive of Pushchan over the mixing primitive of PaV. Demirel *et al.* [148] recently aimed to improve PaV to provide everlasting privacy by using a commitment scheme and zero knowledge proof in the ballot generation anonymization processes. They resorted to specific legal and organizational procedures to protect the privacy of voters.

2.2.2.4 Others

Adida and Rivest [149] proposed the "Scratch&Vote" using scratch-off cards to provide receipt freeness and verifiability at the polling place. Their scheme published votes in

encrypted form, and was, therefore, only computationally private. Another approach to a privacy preserving voting system was Farnel [150], [151] which was not widely accepted due to its complex vote casting procedure. When a voter casts her vote in the Farnel box, a subset of its content is scanned and, accordingly, a receipt is generated by spinning the box. Although conceptually similar to floating receipt mechanism, the advanced hardware dependency remains a limitation.

Among the other techniques [152]-[154] were designed for remote or internet based voting; still considered unrealistic considering the physical limitations at the vote casting point. The early-stage development of electronic voting can be reviewed by interested readers in [162]-[171].

2.3 Conclusion

This chapter has presented the fundamental research strategies and concurrent works on privacy in two relevant contexts of people centric applications, i.e., participatory sensing and electronic voting. Either application is yet to be widely accepted since all the issues relating to privacy are not satisfactorily solved. Hence, the following challenges need to be addressed:

- Traditional approaches add some uncertainty to location or other information in protecting the inferable privacy of an associated observer/reporter. This uncertainty is likely to degrade the quality of the data at the destination and, thus, destroy the ultimate purpose of many applications in the context of PSS.
- To protect the location privacy of observers from remote places, geographic proximity based spatial cloaks are not feasible. Again, the efficiency of anonymization approaches is dependent on the quality of these cloaks. The degree of anonymity also has to be considered. Since PSS is a voluntary participation based technology involving many tiny, power-constrained devices, computational efficiency is desired for any applied anonymization approach.
- In the context of voting, some works emphasize the privacy of voters and coercion resistance or vote trading without ensuring end-to-end verifiability. Other works focus more on verifiability, leaving a loophole in the protection of voter privacy. The few works that have emphasized both these issues simultaneously rely either on complex hardware or unrealistically assume a trusted entity or voter capability.

Chapter 3

3 Subset Coding and Joint Decoding

Inspired by the working principle of PSS, we chose it as the application scenario to test our hypothesis that collective association can automatically preserve privacy without compromising data integrity. The only pre requisition to test this hypothesis is that multiple observers can report about the similar set of individual entities which essentially prevails in the system of participatory sensing. The huge potentiality of PSS to provide a cost-effective alternative to traditional Wireless Sensor Networks (WSNs) inspires researchers to investigate new application domains for this technology. Consequently, a wide variety of application scenarios were identified where the system architecture and role of system entities may vary. In order to protect the inferable location privacy risks, the working principle of different system entities and potential adversary abilities need to be explored. This chapter first discusses and then formalizes the comprehensive system architecture for PSS and then introduces the additional entities required in our proposed location privacy preserving mechanism. It also presents the basics of the proposed subset-coding framework, which (as outlined in Chapter 1) is the corner-stone of different location privacy schemes presented in this dissertation, along with the comprehensive privacy risk analysis, and mitigation strategies.

In this chapter, we first discuss the architecture of a PSS with commonly available entities and then their mode of communication in Section 3.2. The concept of subset-coding technique is presented next in Section 3.3 with some definitions of the terms we use throughout the thesis. We then present a comprehensive adversary model, identify their different malicious capabilities in our proposed architecture and present design strategies to mitigate their attacks to the breach of a participant's privacy in Section 3.4. The risks on privacy of observers when our proposed technique is used are thoroughly analysed in

Section 3.5 followed by extensive simulation in Section 3.6. Finally, some concluding remarks are made in Section 3.63.7.

3.1 Introduction

PSS provides common people with a platform to sense, collect, analyse, and share local information or knowledge for their own benefit. Smart-phones equipped with high precision localization capability, camera or other ad-hoc sensing devices mounted on vehicles are used to record objects/events of interest by the people in course of their daily life. The captured data are sent to specific servers via some lightweight inexpensive wireless communication networks. The collective reports from a large number of participating users help the server generate useful information and reply to the queries of the user on-demand.

In the context of PSS, the privacy of participants who are willing to share their information should be respected, especially when information travels across open wireless networks. Data privacy is not a concern as the reported data is meant to be shared among the community. However, the apparently insensitive information transmitted in plaintext through this lightweight infrastructure can be eavesdropped by an adversary and this information can be used to infer some sensitive information which threatens the location privacy of the observer. Participants are required to provide their identifications while

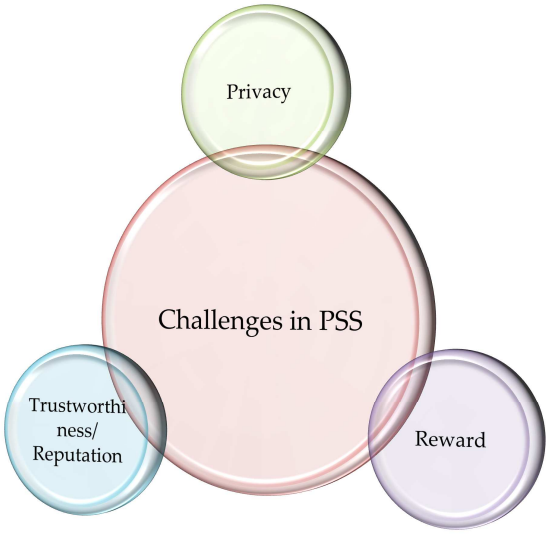


Figure 3.1: The problems to be addressed simultaneously in PSS.

reporting (i) to facilitate building up a reputation scheme that is required to maintain reliable and trustworthy data and (ii) to allow the system develop a reward-based service provision e.g., the more someone contributes the higher will be the quality of data received from the server. Consequently, reports to the server cannot be sent anonymously as hiding the data ownership straight-away cannot provide a complete solution in this context. This multi-dimensional problem is depicted with Figure 3.1.

It would be helpful to understand the significance of this problem if we discuss some application scenarios of PSS in brief.

- PetrolWatch [2]: A network of mobile cameras that automatically collects fuel prices and serves users by reporting on the cheapest fuel station in the user's locality.
- Safe cities [6]: Uses mobile devices to collect and share the safety level of the location such as unreported crimes, or web based applications for visualizing unsafe areas.

Recently, many other consumer price information sharing applications using PSS [3]-[5] have been proposed. Throughout this thesis, we use the example scenario of *PetrolWatch* [2] to present different ideas on the architecture and techniques of privacy preservation. For generic consumer price information sharing applications, both the location privacy and association with products may be revealed from the reports of observers. Although we discuss the case of location privacy in this work, our proposed anonymization technique is readily applicable to protect product-association privacy as well.

The existing location privacy protection mechanisms [36]-[116], where inferable location information is transmitted with some anonymity or by adding Gaussian noise or at reduced precision, cannot be used where the destination expects complete data integrity at individual level. Unless sufficient data quality/accuracy is achieved, users will not be encouraged to use the system. For example, if *PetrolWatch* cannot assist drivers to find the cheapest fuel station in the neighbourhood and recommends an expensive one instead, the reputation of this application will be at stake. Hence, data quality/integrity and privacy needs to be protected concomitantly. Therefore, a privacy-preserving data communication technique is needed such that each observation of the attribute of a Point of Interest (POI) by a participant can be transmitted with sufficient POI-level anonymity such that the data collector can eventually de-anonymize individual data, i.e., associate the attribute with the

correct POI. As long as an adversary is unable to intercept a reasonably high number of transmissions from the participants, any de-anonymization attempt to infer the attribute-POI association remains sufficiently ambiguous.

Our solution to address all these significantly challenging problems are based on a novel technique called subset-coding. The privacy of the participants who reports an observation about a visited POI is ensured by *k*-anonymization. The concept of *k*-anonymity states that an observation is *k*-anonymous if the observed POI is indistinguishable from $k-1$ other POIs. The feasibility of the technique relies on an efficient joint de-anonymization technique to obtain sufficient data quality at the desired end. The anonymization and de-anonymization techniques developed are presented in subsequent chapters. We assume no secure communication channels or the possibility of adversaries to collude while designing robust system architecture with viable protocols to safeguard against all types of adversaries. Most of the existing privacy-preserving techniques as already presented in detail in Chapter 2, cannot be a viable solution so long as high data quality is concerned.

3.2 System Overview

In Section 3.2.1 the system entities of our proposed scheme are presented along with some relevant definitions. The system model and its functionalities are then described in Section 3.2.2.

3.2.1 System Entities

The basic entities of our proposed participatory sensing system are described below.

- **Mobile Nodes (MNs):** MNs are the users that collectively sense attributes of N POIs and report them.
- **Application Server (ApS):** The ApS is the server that receives anonymized reports from users, or in other words MNs, and based on these reported information, it makes the service available for the users such as replies to user query or informs users about the attributes of POIs, e.g., the cheapest fuel station in their vicinity (considering *PetrolWatch*). It has a decoding application that decodes the attributes of the POIs from anonymized reports received from the MNs.

- **Anonymization Server (AS):** A third party AS is used to achieve desired POI-level anonymity while remaining transparent to the ApS. Most of the existing techniques rely on an AS, a trusted third party, to perform the anonymization centrally.

In our proposed scheme, MNs send plain and simple Observation Reports (OR) to AS. This is simply a POI-attribute pair related to an observed POI. For POI i having attribute a_i an OR o_i will assume the form $[i, a_i]$ and is sent as plain text. Upon receiving an OR our AS k-anonymizes it using a greedy heuristic such that the more anonymized reports are received by ApS, the higher is its ability to associate POIs to correct attributes. The anonymized report is then relayed to the observer. Finally, the observer MN sends the anonymized report to the ApS. Although the anonymization can be performed independently by selecting additional $k - 1$ POIs at random, this random anonymization can hardly influence the data integrity to reach the target. Moreover, most of the existing techniques rely on a trusted third party AS to achieve homogeneity among selected POIs so that any of them remains equally likely to thwart any decoding attempt by an adversary. At this point we introduce our AS to perform the anonymization centrally, i.e., a directional anonymization such that target data integrity can be reached efficiently with received collective data. Note that $k \in \{2, \dots, N-1\}$ as $k = 1$ offers no anonymity and $k = N$ makes decoding impossible. Let us term the k-anonymized form of an OR as *Anonymized Rule* (AR) defined as follows.

Definition 3.1 (Anonymized Rule): *An Anonymized Rule (AR) for OR* $[i, a_i]$ *is expressed as* $AR_i \equiv \{i_1, i_2, \dots, i_k\}: a_i$ *where* $\{i_1, i_2, \dots, i_k\} \subset \{1, \dots, N\} \wedge i \in \{i_1, i_2, \dots, i_k\}$ *are selected by AS using a greedy approach.*

For example, when 3-anonymity is desired, \$10 price observed for POI 1 may use AR $\{1, 2, 3\}: \$10$ to anonymize POI 1 with POI 2 and POI 3. The detail of the greedy anonymization approaches are discussed in Chapters. 4 and 5. After receiving an AR, the observer sends a Report towards ApS (RApS) defined as follows.

Definition 3.2 (Report towards ApS): *A Report towards ApS (RApS) may be expressed as [User_id, AR, time_of_observation].*

An observation of POI 1 by MN having user id $u1$ observed at time t then may assume the form towards ApS as $[u1, \{1, 2, 3\}: \$10, t]$.

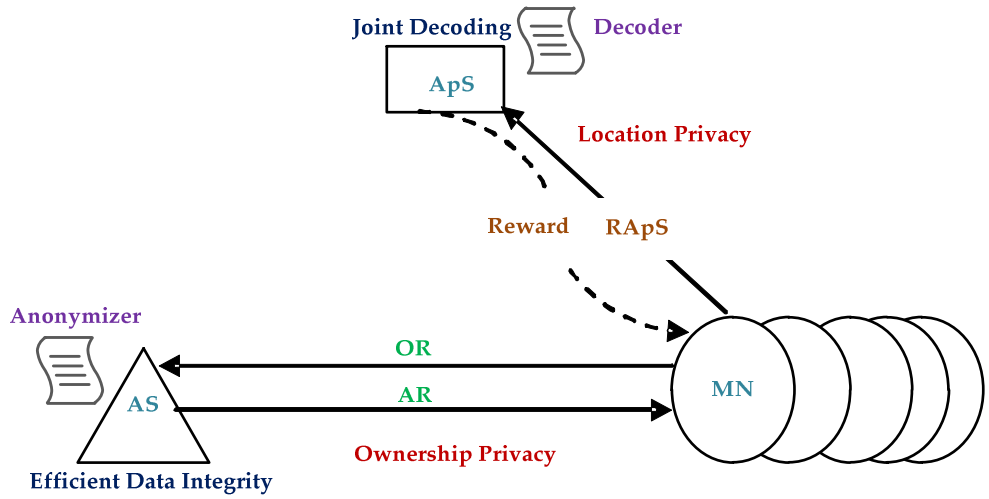


Figure 3.2: Conceptual Diagram of our proposed PSS.

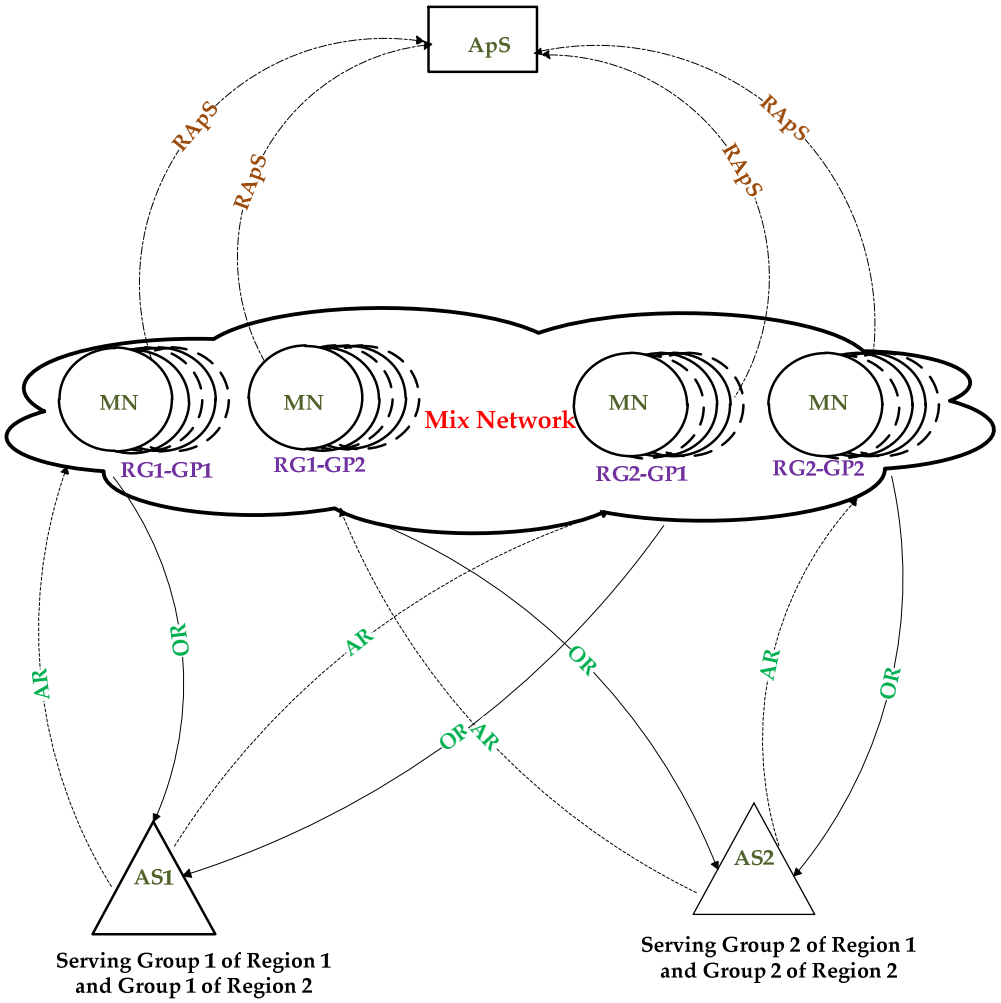


Figure 3.3: Detail schematic of proposed PSS.

3.2.2 System Model

Figure 3.2 presents these entities along with their properties and of the communication medium. Note that before anonymization, any observation directly sent from the user can pose a threat to the user's privacy straightaway in case of eavesdropping. Hence, we propose that no direct communication would take place between the observer MN and AS to safeguard against possible eavesdropping of unanonymized ORs. Moreover, since AS neither provides any reward nor cares about the reputation of the source, user id is not needed to be attached during the communication between an MN and AS.

At this point, we introduce a friend network to deal with this issue of ownership privacy. That is, the OR is relayed through a mix network of random length before delivering to AS. As the length of the chain is random, it is impossible to guess the originator. Here the challenge is to make this ownership privacy maintained communication bidirectional. In one way the MN sends OR to AS, while on the other way the AS sends the suggestion back to the originator. Introducing the concept of mix network may develop a particular type of adversary who may join as a member of this mix network. To handle adversaries we propose dividing the overall user domain into regions, each having a number of groups in it served by multiple ASs. The details of various possible adversaries and strategies against them are discussed in Section 3.4.

For the sake of completeness, the detail schematic of the proposed conceptual diagram is given in Figure 3.3. Here, AS1 deals with the ORs from group 1 of region 1 and group 1 of region 2 whereas AS2 serves those from group 2 of region 1 and group 2 of region 2. The mixer of messages for the friend network is shown by the cloud which implies that communication with the corresponding AS takes place through the cloud, that is, the friend network. The solid MN indicates the physical MN located in a particular instance whereas the dotted one indicates that the same MN can also report from a different group in a different instance. An individual MN can report from different groups of different regions and that is represented by the dotted version of that solid MN. Hence the design principal states that,

- All groups of a physical region should not be served by the same AS. This is done to handle rival-minded AS such that it fails to get the full picture as presented in Section 3.4.1.
- At the beginning, when an MN registers as a user in the system, it receives different user id to be used while reporting from different corresponding groups.
- As the user ids against a single user are issued by the ApS at the registration, it can distinguish reports from a group from that knowledge.
- A friend network is not bounded by this grouping of regions. Hence, the adversary fails to speculate which users belong to the same region/group.
- At the registration, each MN is provided with a number of (network) friends randomly identified by their network id. This id is different from user id as used in different groups of regions.

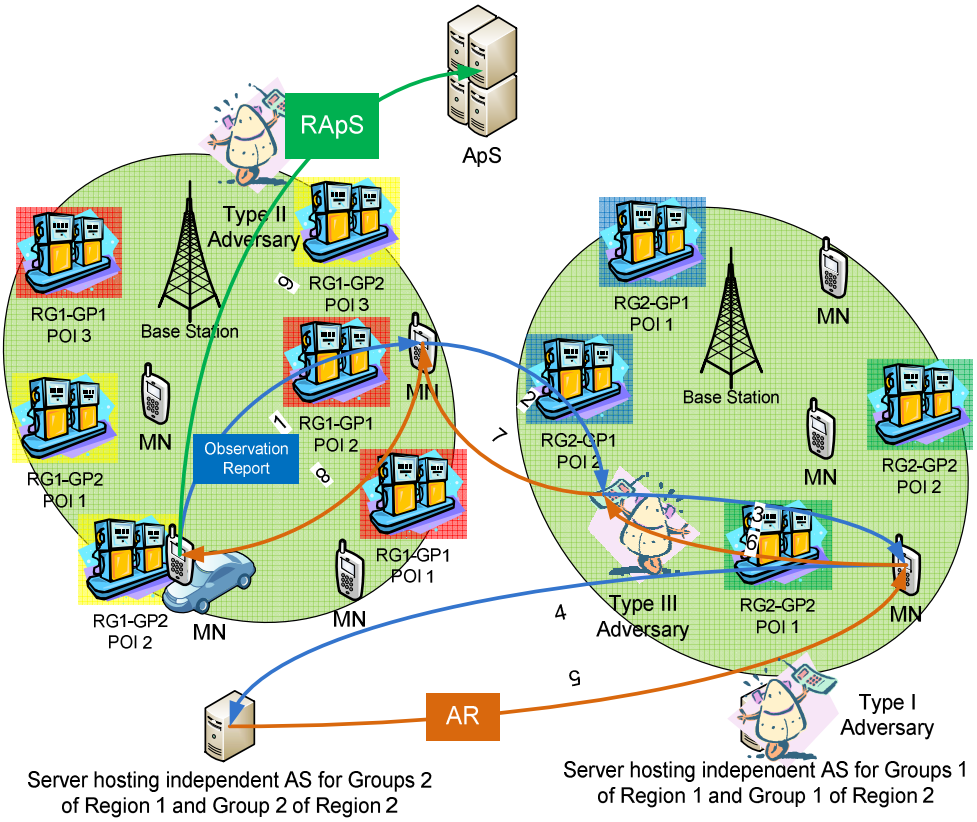


Figure 3.4: Entities and information flow of a typical system scenario of the proposed PSS.

Figure 3.4 presents a practical example of our proposed system scenario along with a typical information flow sequence (numbered 1 to 9) from observing MN to the ApS in the *PetrolWatch* context. It also consists of three adversaries that are discussed in detail in Section 3.4.1. Here, the MN near POI 2 of group 2 of region 1 originates the OR and sends it to the corresponding AS using its friend network in the flow sequence numbered from 1 to 4. Then, AS sends back the corresponding AR following the same route of the friend network in reverse order in the flow sequence numbered from 5 to 8. Finally in sequence 9, the RApS is uploaded to the ApS from the originator MN.

The computational complexity of anonymization as well as de-anonymization is significantly higher when the attributes of different POIs are not unique, which is naturally occurring as POIs are assumed to be non-communicating. However, the non-unique scenario can be easily transformed to the unique scenario with the assistance of the AS as explained later. Therefore, the attributes of N POIs are almost everywhere in this thesis (except for the risk analysis in Section 3.5.3) assumed unique without any loss of generality.

3.3 Basic Concept

In Section 3.3.1 we present the concept of subset-coding that is the basis of our proposed anonymization techniques, followed by that of the proposed joint decoding in Section 3.3.2. The basic concept is explained in the context of *PetrolWatch* [2] where each participant independently reports the observed petrol price.

3.3.1 Subset-Coding

In our work we aim to preserve the location privacy of the MN via k-anonymization of the observation. We assume there are N numbers of POIs whose prices are uniquely defined. Our AS receives actual price and corresponding POI from the MN. Then to make a report with k-anonymity, the actual POI can be anonymized with any $k-1$ out of the remaining $N-1$ POIs. Let us consider a *PetrolWatch* scenario with $N=4$ petrol pumps of ids 1, 2, 3, and 4. For a desired anonymity level $k=2$, an OR [1,10] can be anonymized using any of the subsets {1,2} or {1,3} or {1,4}. Thus, the POI attribute is reported making the POI anonymous among $k-1$ other POIs and consequently individual location privacy is maintained. This selection of $k-1$ other POIs can be random or can follow a particular guideline. The detail of how the subsets are selected in anonymized rules is elaborated on in the next chapters.

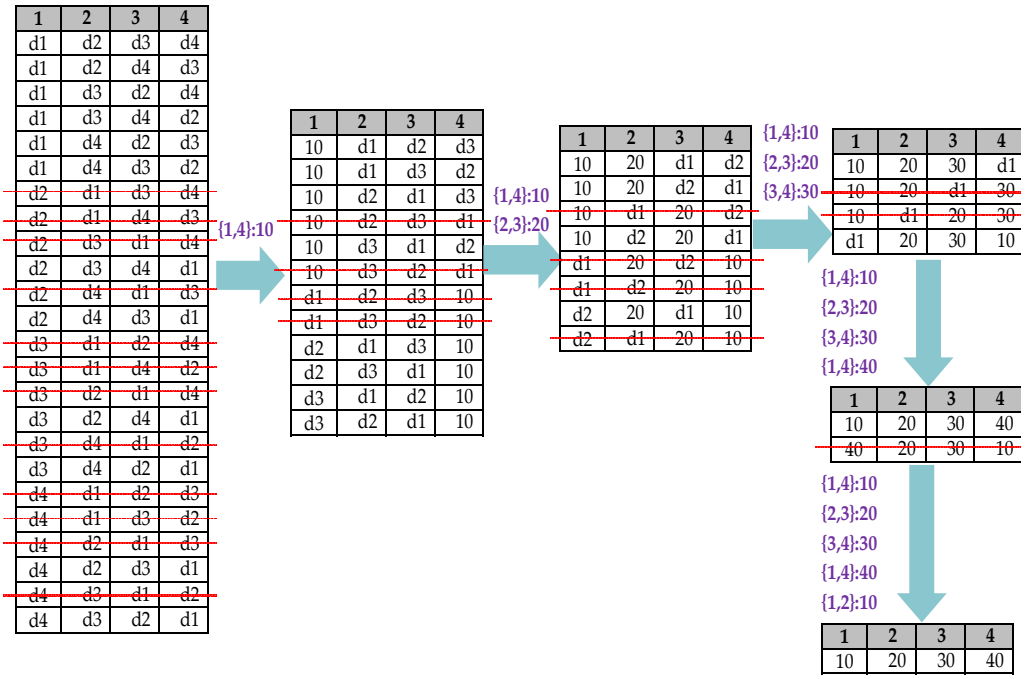


Figure 3.5: Conceptual depiction of joint decoding.

3.3.2 Joint Decoding

For the example given above let us assume that in response to five ORs $[1, 10]$, $[2, 20]$, $[3, 30]$, $[4, \$40]$, and $[1, 10]$, the AS generated the following $k(= 2)$-anonymous ARs $\{1,4\}: 10$, $\{2,3\}: 20$, $\{3,4\}: 30$, $\{1,4\}: 40$, and $\{1,2\}: 10$, respectively as shown in the example in Figure 3.5. Note that "$" is omitted for the sake of brevity. The generation algorithm for these ARs, or in other words the anonymization technique, may be random in the simplest case or designed based on some heuristics. For the time being, we ignore this issue and will discuss it in subsequent chapters. At the very beginning, the decoding process starts with all possible mappings between POIs and their dummy prices that are identified with leading d. Each AR rules out some of these mapping possibilities. For example, when the first AR $\{1,4\}: 10$ arrives at the decoder of ApS, it removes the tuples that associate either POI 2 or POI 3 with d1 and then replaces d1 with 10. With the gradual arrival of the ARs, the dummy prices are replaced by the coming attributes. Finally, when only one possibility is left, full decodability is achieved, as realized in Figure 3.5.

Definition 3.3 (Decodability): *A particular outcome of an anonymization scheme satisfies D-decodability iff D or more POIs can be associated to their correct attributes. N-decodability is also referred to as full decodability.*

Elaborating the example in Figure 3.5, the AR $\{A, D\}: 2$ implies that either A or D has price \$2. Hence, the permutations (throughout the thesis we have used another term tuples interchangeably) indicating either B or C has price \$2 are removed. In this way, 12 possible mappings are ruled out. The remaining $4! - 12 = 12$ possibilities are now checked for subsequent ARs and some possibilities are similarly ruled out for each of them. Finally, only the actual association remains as conforming to all the ARs and full decidability is achieved. Note that if no tuple survives in this approach, it indicates that either some false data has been reported or the price of one or more POI has changed. The impact of such cases on our system will be discussed in Chapters 4 and 5.

In the next section we analyse the location privacy risk of participating users in our anticipated PSS architecture when their anonymization is done using subset-coding.

3.4 Adversary Models and Risk Mitigation Strategies

The focus of a malicious third party is to reveal the location information included in the user reports. From these eavesdropped messages, it tries to infer POI-attribute association using the same decoding approach of the ApS and, thus, reveal the location of the observer near that POI. It is natural that the adversary has close access to victim in real world. Therefore, the victim's user id is known. We first explore adversaries with various capabilities in Section 3.4.1 and then discuss strategies to counter them in Section 3.4.2.

3.4.1 Different Types of Adversaries

The adversaries that attempt to reveal the location information from the user reports working through the PSS communication interception may cause five potential privacy risks as discussed below:

- **Type I Adversary (Compromising AS):** An adversary compromising the AS can access the unanonymized ORs but they have no user id attached. Moreover, after introducing the mix network, the originator's identity cannot be traced back. Moreover, as it is a requisite from the security surveillance that the public communication should remain plain-text, we may assume that the service of the AS is provided by some trusted entity e.g., a government agency.

- **Type II Adversary (Eavesdropping RApS):** Residing near the ApS, adversary can receive quite a good number of RApS. Having equipped them with the same decoding application as ApS, it is likely to infer significant user-POI association.
- **Type III Adversary (Capturing OR from mix network):** It is highly likely that the adversary is a member of the friend network. Thus, the adversary will receive a certain amount of messages.
- **Type IV Adversary (Compromising ApS):** If an adversary is able to compromise someone having access to ApS, i.e., an employee. This may be possible only for a short duration, then it may have access to ApS data collection.
- **Type V Adversary (Compromising POIs):** Some eavesdropping devices may be mounted in the vicinity of the POIs.

Based on the above discussion, we concentrate on the first three cases namely Type I, Type II, and Type III depending on what messages they eavesdrop. Because in the case of Type IV, the basic security mechanism of ApS such as an audit trail and access logs will not allow its compromised employee to continue the undue activity for long. Hence, the data captured in this way is limited in size and is likely to reveal little information. Considering Type V, we propose that the MNs will transmit the ORs only after going outside some pre-defined distance e.g., the coverage of the POI's base station so that the observation and transmission will not occur from the same place. Hence, from this point we focus on addressing the other three adversary threats.

3.4.2 Risk Mitigation Strategies

In this section, we discuss specific strategies to mitigate risks against the three types of adversaries identified above.

3.4.2.1 Strategies against Type I Adversaries

Type I adversary is a rival-minded AS who aims to challenge ApS and provide service on its own. The compromised AS in our system cannot pose a threat to location privacy, but has the actual information about POI-attribute association from the ORs accumulated. Thus, it can throw a service based challenge to our ApS.

We propose to guard against Type I adversaries by dividing POIs in a region into several groups, each served by a dedicated AS. In Figure 3.4, two groups of POIs are used in

each region covered by a mobile base station. Now, in case an AS gets rival-minded to the ApS, it can never guarantee complete information about the whole scenario. Moreover, all the ASs of the system are compromised and collaborating with each other to give complete service is a too extreme case to assume.

3.4.2.2 Strategies against Type II Adversaries

The above methodology of dividing a region into several groups also serves the purpose of handling Type II adversaries. We propose that a user is allowed to either sense POIs from only one group or use separate user id for each group. As the adversary is unable to distinguish RApSs from different groups of users, any attempt to associate POIs to correct attribute will fail miserably unless in a rare situation the attributes of all correspondingly numbered POIs in all groups are same.

ApS, however, can distinguish RApSs from different groups properly from the user id and, hence, can associate POIs to correct attribute. Allowing a user to report on the POIs of only one regional group using one user id is quite practical in the context of consumer price sharing since users would report on their own locality frequently. If a user frequently travels to another area, she may register with another user id for that area. In contrast, dividing the region into several groups is also good for keeping the computational complexity at a user-friendly level.

Moreover, the physical limitations of standing for a long period in a monitored public place can also play a defensive role against this adversary. Hence they are assumed not to be able to eavesdrop sufficient number of RApSs to decode the association of POIs to their actual attributes with reasonably high accuracy.

3.4.2.3 Strategies against Type III Adversaries

Type III adversary is actually a compromised network-friend that emerges as we introduce the mix network into the scenario. They capture the plain-text ORs and, thus, learn the correct attributes of one or more POIs. Although Type III alone cannot pose any risk to location privacy of the target victim due to absence of user id in the OR, she may collude with a Type II adversary to improve the decoding accuracy as follows. The Type III adversary may gain the correct attribute of one or more POIs by successfully intercepting some OR passed through the mix network. In the colluding scenario, this information is shared with the Type II adversary who can then either reduce the degree of anonymity (k)

for some AR by eliminating possible POIs in the list and/or improve decoding accuracy by effectively reducing the impossible set of POIs for each reported attribute.

Risk analysis of location privacy by such a colluding pair of adversaries is a challenging task. We have performed thorough analysis as presented in the next section.

3.5 Risk Analysis

In this section, we discuss the risk evoked from introducing the mix network as they face different types of adversary attacks. First, analytical analysis on the probability of intercepting necessary OR to learn attributes of $0 \leq \tau \leq N$ POIs by a Type II adversary are shown in Section 3.5.1. Then Section 3.5.2 quantifies the risk of disclosing the whereabouts of the targeted victim when the adversaries collude and this is done for the strictest scenario where the POI attributes are considered unique. Finally, in Section 3.5.3**,** the analysis is generalised by modelling the probability of attribute uniqueness and, then, extending the risk model using both interception and uniqueness probabilities.

3.5.1 Interception Probability

In the mix network scheme, OR flows unencrypted through a certain number of members in the mix network domain towards the AS. By definition a Type III adversary is likely to be a member of this network who may learn the attributes of some POIs through active participation within the network. The key factor behind the risk here is that whether or not the adversary is able to learn that POI's (where its target is located) current attribute from its received data in random data flow through the mix network. To model this ability of adversary and its incurred risk on the whole system, we performed the following analysis.

Let n be the number of total registered users in the system and for the sake of simplicity let us assume that each user preselects F network friends either at random or with the help of ApS. Let us also assume that h is minimally set by the system such that any of the n users in the system remains equally likely to be the observer of an intercepted OR, i.e., $F^{hp-1} < n \leq F^{hp}$. Also note that to provide user anonymity the expected number of hops hp must be greater than 1 such that the immediate previous hop may avoid the risk of being exposed as the originator. So, we may conclude,

$$hp = \max(2, \lceil \log_F n \rceil) = \max\left(2, \left\lceil \frac{\ln n}{\ln F} \right\rceil\right). \tag{3.1}$$

Let P_τ denote the *interception probability* that the adversary's intercepted reports contain attributes of $0 \leq \tau \leq N$ distinct POIs. Using a simple simulation of the mix network the probability density function (pdf) of P_τ is estimated and the average results are reported in Section 3.6.2.

3.5.2 Risk of Location Privacy with Unique Attributes

Let the attribute of POI 1 at time t be \$10. Suppose the Type II adversary of the colluding pair could somehow eavesdrop an RApS $[u1, \{1, 3, 6\}: \$10, t]$ from the target victim, $u1$. We are interested in estimating the probability of location privacy risk P_{risk} that the Type III adversary would be able to intercept the relevant OR $[1: \$10]$ passing through the mix network. When this information is matched against the intercepted RApS, the location of the victim at time t will no longer be anonymized if attributes of POIs are unique. Although a very unlikely scenario, assuming unique attributes of POIs allows us to develop the analysis in a simple way first and then give us insight to extend it for the generalised non-unique scenarios in the next section.

We can estimate the maximum risk from the following average-case analysis. As each OR is passed on to hp users on average and ωn observations are made on average during the temporal window, in total ωnhp reports are seen by n users during this period, i.e., each user is expected to receive on average ωhp reports. As the adversary is one of the users and assuming that the ORs received by the adversary contains no duplication, maximum risk can be estimated as

$$P_{\text{risk}}^{\max} = \frac{\omega hp}{N} = \frac{\omega}{N}\left\lceil \frac{\ln n}{\ln F} \right\rceil. \tag{3.2}$$

As the likelihood of intercepting identical ORs by the adversary increases with ω, the actual (average) risk will be smaller than $P_{\text{risk}}^{\max}$ for large ω. If the Type III adversary intercepts the attributes of $0 \leq \tau \leq N$ POIs, then the chance of having the target attribute (\$10 in the above example) among these τ unique attributes is τ/N. Hence, we can calculate the expected risk as

$$P_{\text{risk}} = \sum_{\tau=1}^{N} \frac{\tau}{N} P_\tau. \tag{3.3}$$

Note that, the parameters N, n, and ω are fixed for a specific system. The number of POIs N cannot be lower than the desired level of anonymity while it cannot be higher than

the number of POIs in a geographical proximity. The typical value of N is in the range [4,8]. Population size n depends on the popularity of the scheme; while ω is governed by the socio-economic and cultural behaviour of the users. By attenuating the only variable parameter F, it is possible to design the mix network such that the maximum risk probability is within a user-defined threshold R, i.e.,

$$P_{\text{risk}}^{\max} \leq R. \tag{3.4}$$

Let F_R denote the average number of network friends per user such that relation 3.4 holds. We may then conclude

$$\left\lceil \frac{\ln n}{\ln F_R} \right\rceil \leq \frac{NR}{\omega} \Rightarrow \frac{\ln n}{\ln F_R} \leq \left\lfloor \frac{NR}{\omega} \right\rfloor \Rightarrow F_R \geq \left\lceil n^{1/\left\lfloor \frac{NR}{\omega} \right\rfloor} \right\rceil. \tag{3.5}$$

Corresponding simulation result with discussion is given in Section 3.6.3.

3.5.3 Risk of Location Privacy with Non-unique Attributes

So far we have assumed that attributes of N POIs are unique. In real-world scenarios, however, this assumption is unrealistic as a number of POIs may share the same attribute among them. This is because, in any time period, attributes are drawn from a small domain or they are highly correlated. Therefore, it is important to develop a suitable model to estimate the probability of uniqueness of attributes. Without any loss of generality, we assume that attributes are drawn from the current values of N POIs in the estimation of this probability.

Let P_η denote the uniqueness probability that the attributes of $\eta \in \{1, \ldots, N-2, N\}$ POIs are unique. Note that $\eta = N - 1$ is an invalid statement. In order to estimate this probability, we need to develop a systematic way of identifying all possible templates of how the

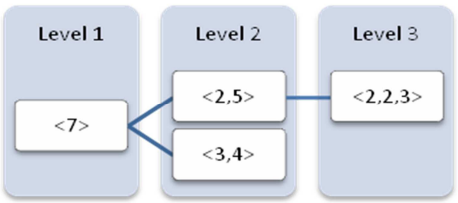


Figure 3.6: The general tree representing all possible templates of grouping 7 POIs into subsets of size two or more.

remaining $N-\eta$ POIs can be grouped into subsets of size two or more. Let us first consider a simple example where of all the $N=8$ POIs only $\eta=1$ will have a unique attribute. Then the remaining $N-\eta=7$ POIs may have non unique attributes according to the possible templates as shown using a general tree in Figure 3.6. Note that as they exhibit non-unique attribute, the possible templates of grouping these 7 POIs into subsets starts from a minimum size of two, then more. Careful observation of the tree reveals that each node, representing a grouping where the subsets are ordered by size, may lead to other groupings in the next level by dividing the largest subset in the group into two subsets of size no less than the size of the preceding subset.

Let us now calculate the number of ways in which these groupings can be achieved by assigning attributes to N POIs. Again, let us first consider a specific grouping $\langle 2,2,3\rangle$ from the above example. The unique attribute ($\eta=1$) can be drawn from N possible values in $\binom{N}{1}$ ways. Two attributes with same subset cardinality of 2 can be drawn from the remaining $N-1$ possible values in $\binom{N-1}{2}$ ways. The remaining attribute can be drawn from the remaining $N-(1+2)$ possible values in $\binom{N-3}{1}$ ways. These four selected attributes can then be assigned to N POIs in $\frac{N!}{2!2!3!}$ ways.

Considering total N^N possible ways, we can generalise the above observation to express *group probability* of the grouping $\langle s_1,\dots,s_j\rangle$ as

$$P_G\big(\eta,\langle s_1,\dots,s_j\rangle\big)=\frac{\binom{N}{\eta}}{N^N}\times\prod_{l=1}^{\left|\mathcal{F}_{\langle s_1,\dots,s_j\rangle}\right|}\binom{N-\eta-\sum_{q=1}^{l-1}\mathcal{F}_{\langle s_1,\dots,s_j\rangle}(q)}{\mathcal{F}_{\langle s_1,\dots,s_j\rangle}(l)}\times\frac{N!}{\prod_{l=1}^{j}s_l!} \tag{3.6}$$

where $\mathcal{F}_{\langle s_1,\dots,s_j\rangle}$ denotes an array of frequencies of distinct elements in $\langle s_1,\dots,s_j\rangle$. The tree construction algorithm can now be generalised to calculate the η-uniqueness probability as

$$P_\eta=\begin{cases}P'_\eta(\langle N-\eta\rangle), & \eta\le N-2;\\ P'_\eta(\langle\,\rangle), & \eta=N;\\ 0, & \text{otherwise}\end{cases} \tag{3.7}$$

where

$$P'_\eta\big(\langle s_1,\dots,s_j\rangle\big)=P_G\big(\eta,\langle s_1,\dots,s_j\rangle\big)+\sum_{k=max\left(2,s_{j-1}\right)}^{\left\lfloor\frac{s_j}{2}\right\rfloor}P'_\eta\big(\langle s_1,\dots,k,s_j-k\rangle\big). \tag{3.8}$$

Let us now formulate location privacy risk when attributes of η POIs are unique and the Type III adversary has already intercepted attributes of τ POIs. If the colluding Type II adversary eavesdrops a RApS from the target victim containing attribute \$10, we are interested in estimating the risk that the associated POI, i.e., the whereabouts of the victim, is exposed. In short, risk arises for the situation of the desired attribute's being intercepted and being unique as well. Let S_U be the set of η unique attributes and S_I be the set of τ intercepted attributes. Risk to location privacy exists iff $\$10 \in S_U$ as well as $\$10 \in S_I$. The probability of $\$10 \in S_U$ is $\frac{\eta}{N}$ and that of $\$10 \in S_I$ is $\frac{\tau}{N}$. Considering that the colluding adversaries working in disjoint domain, Type II intercepting RApSs and Type III intercepting ORs, these two probability can also be assumed disjoint to estimate the risk of location privacy as the product of these two probabilities, i.e., $\frac{\eta\tau}{N^2}$.

We can now estimate the expected risk for non-unique scenarios as

$$P_{risk} = \sum_{\tau=1}^{N} \sum_{\eta=1}^{N} \frac{\eta\tau}{N^2} P_\eta P_\tau. \tag{3.9}$$

Note that in the unique scenario, domain of η is a single value N and $P_\eta = 1$. Substituting these equalities in 3.9 transforms it into 3.3.

3.6 Results and Discussion

In this section, we present simulation results to validate a number of key findings throughout this chapter. The simulation setup is described in Section 3.6.1 followed by various simulation results to find interception probability in Section 3.6.2. Then in Section 3.6.3 a simulation guideline to control maximum risk probability is shown, followed by a presentation of the unique attribute probability distribution function in Section 3.6.4. Finally, a comparison of risk probability when attributes are unique and non-unique is presented in Section 3.6.5.

3.6.1 Simulation Setup

To improve simulation accuracy each simulation was repeated 1000 times and for each setup, each user randomly selected F network friends out of $n-1$ other users and randomly selected ωn users then observe POIs at random and send ORs to the AS using a chain of network friends of expected length $\left\lceil \frac{\ln n}{\ln F} \right\rceil$ as estimated in (3.1). We have considered $n \in$ {100,10000}, representing small and large PSS, respectively and number of network friends

used by each user is carefully controlled so that the comparable average number of hops hp can be achieved for both populations using the minimum number of network friends. For $n = 100$, we have used $F \in \{3,10\}$ and for $n = 10000$, we have used $F \in \{7,100\}$ to effectively guarantee $hp \in \{5,2\}$ for both populations. To simulate a real-world environment, we have introduced observation rate ω such that at any observation period, ωn random users report sensed data.

3.6.2 Interception Probability Distribution

In our simulation, each user maintains a list of POIs from which it could find attributes from

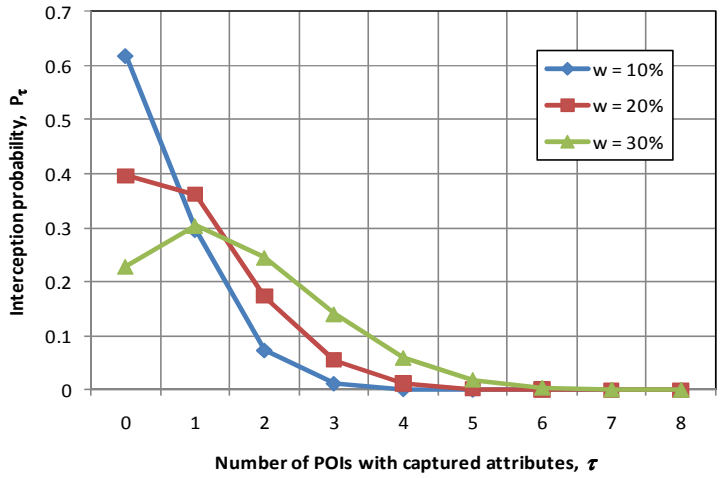


Figure 3.7: Interception probability distribution for $n = 10000$, $N = 8$, and $F = 7$ at $\omega \in \{0.1,0.2,0.3\}$.

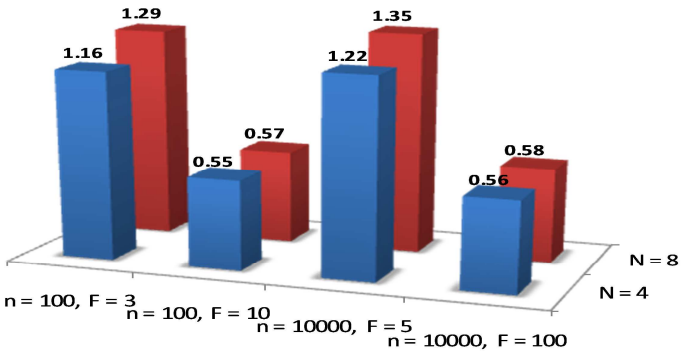


Figure 3.8: Expected number of intercepted POIs at $\omega = 0.3$.

Table 3.1: Lower Bound on Number of Network Friends so that Maximum Risk Probability ≤ 0.1, i.e., $F_{0.1}$ when $N = 8$.

	No of users, n				
ω	***100***	***1,000***	***10,000***	***100,000***	***1,000,000***
0.1	2	3	4	5	6
0.2	4	6	10	18	32
0.3	10	32	100	317	1000

messages passed through it, which eventually allows the simulation to calculate the pdf of P_τ. In Figure 3.7, we have plotted the interception probability distribution for $n = 10000$, $N = 8$, and $F = 7$ at $\omega \in \{0.1,0.2,0.3\}$. Notice how the distribution profile is shifted to right as observation rate is increased. Ultimately, the expected number of intercepted POIs, $\sum_{\tau=0}^{N} \tau P_\tau$, is increased from 0.48 to 1.58. Note that even at 30% observation rate, which is quite high, compared to a real-world rate, the expected number of intercepted POI just 20% of all POIs.

Figure 3.8 presents the expected number of intercepted POIs by a Type III adversary for different values of n, N, and F at $\omega = 0.3$. Clearly, the number increases with n and/or N and decreases with F when other two parameters remain static. Note that the average number of hops used by the system decreases with F and, hence, the total number of messages potentially intercepted by the adversary is also reduced. We have deliberately used the minimum possible F for both population (10 and 100 for population 100 and 10000, respectively) to operate with the minimum possible expected hop count, i.e., 2 to estimate the minimum possible risk. We have observed that the minimum possible expected number of intercepted POIs is insensitive to n and N. For $N = 8$ and $n = 10000$, attributes of only 7% POIs may be intercepted by an adversary and to achieve this, each user needs to assign merely 1% of the population as network friends.

3.6.3 Attenuating Maximum Risk Probability

Table 3.1 presents the lower bound on $F_{0.1}$ for different n and ω with $N = 8$. Clearly, we need to assign more network friends per user to keep the risk low as the observation rate increases. However, the overhead of increasing F is minimal. The storage requirement of the network friends' user ids is also not an issue with the use of very cheap solid-state memory.

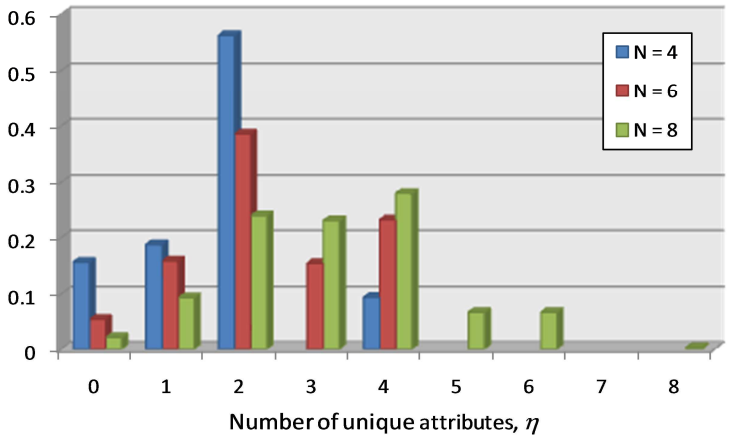
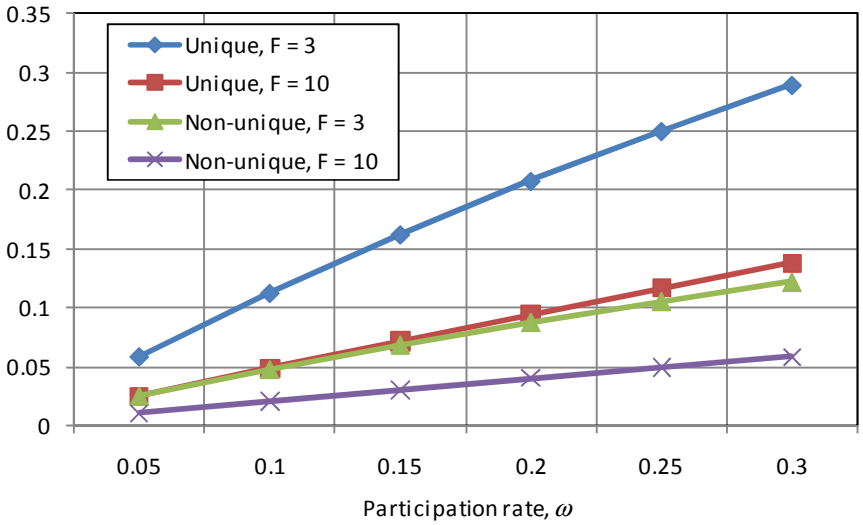


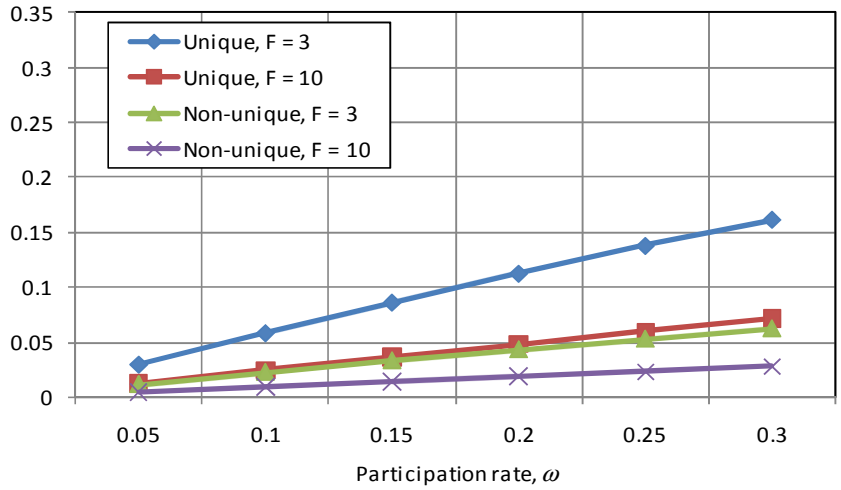
Figure 3.9: Uniqueness probability distribution for $N \in \{4,6,8\}$.

3.6.4 Unique Attribute Probability Distribution

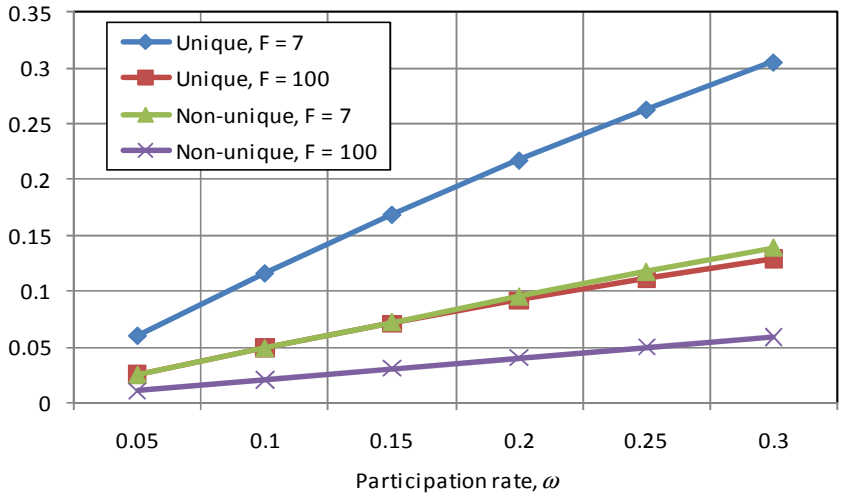
Figure 3.9 plots the unique attribute probability distribution, P_η for $N \in \{4,6,8\}$. Here we see that POIs having non-unique attributes is actually a practical scenario. All the POIs having an unique attributes is a very rare case. When $N = 4$, it is highly probable (55%) that only 2 of them have unique attributes. All the four unique attributes may occur in only 9% cases.



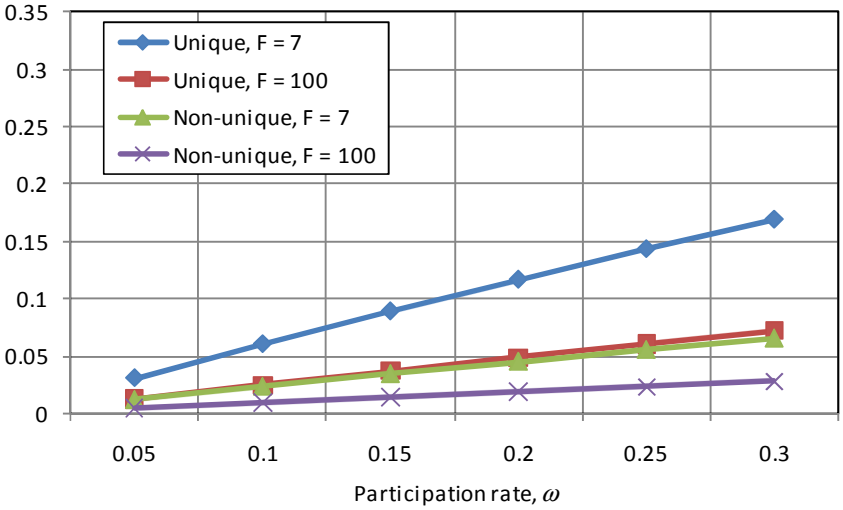
(a) $N = 4$ and $n = 100$



(b) $N = 8$ and $n = 100$



(c) $N = 4$ and $n = 10000$



(d) $N = 8$ and $n = 10000$

Figure 3.10: Location privacy risk P_{risk} for unique and non-unique attributes.

3.6.5 Unique vs Non Unique

Here, we present our results on location privacy risk P_{risk} due to the mix network for both unique and non-unique attribute scenarios. We have considered $N \in \{4,8\}$, as the computational complexity of the decoder to be presented in the next chapter is significantly high for larger N, and $n \in \{100, 10000\}$, representing small and large PSS, respectively. The number of network friends used by each user is minimally set so that the average number of hops for both population sizes remains the same. For $n = 100$, we have considered $F \in \{3,10\}$ and for $n = 10000$, we have used $F \in \{7,100\}$ needing 5 and 2 hops on average, respectively, to provide user anonymity across the entire population. In all cases, we have considered user participation rate ω in the range $[0.05, 0.3]$.

Figure 3.10 presents P_{risk} for four different setups. In all cases, the location privacy risk increases almost linearly with ω. When the number of network friends per user is adjusted so that the mix network operates with a fixed average number of hops, the risk is almost independent of n as observed in Figure 3.10(a) vs Figure 3.10(c) and Figure 3.10(b) vs Figure 3.10(d). In all cases, the risk decreases with N. Moreover, the risk is almost halved when attributes are considered non-unique, which is indeed the real-world scenario. For $N = 4$, location privacy risk is below 6% for the entire range of user participation rate and for $N = 8$, the risk is below 3%. So, we may fairly conclude that the location privacy risk due to the introduction of the mix network to provide user anonymity while communicating with the AS is very small.

3.7 Conclusion

In this chapter, we have presented our proposed system architecture of PSS along with the flow of information among different entities. We have introduced the system entities needed to design a privacy-protection scheme for the users of PSS. The basic concept of subset-coding technique that was developed as the basis of the anonymization techniques to be presented in subsequent chapters is also discussed here. Finally, different potential adversary risks are discussed in a detail analysis of location privacy risk. We have found that location privacy risk in our approach depends on number of POIs, user population size, user participation rate, and the number of network friends used by each user. We have also observed that the risk is almost halved when observed attributes are non-unique and, more

importantly, the risk can be reduced to a very low level by simply expanding the network friend list of each user.

In the next chapter, we present the subset-coding based *k*-anonymization schemes that achieve the desired data quality at the target end even with a reasonably small number of user observations.

Chapter 4

4 Probabilistic Techniques to Achieve Location Privacy and Data Quality

In the previous chapter, the system architecture of PSS was presented along with the concept of a novel subset-coding technique. In this chapter, we develop a greedy *k*-anonymization scheme that works on that architecture using subset-coding. The experimental results presented in the final section of this chapter supports that availing greedy techniques for optimization is sufficient for this particular problem scenario. Being a voluntary system, it is quite likely that the collection of information will not be very extensive. Keeping this in mind, we aim to guide the anonymization scheme in a probabilistic manner such that the decision made by ApS, on the basis of current collection of information, should reflect the actual scenario for the majority time of this service. The primary goal of designing the anonymization scheme is to ensure high data quality from a reasonably small number of observation reports. The relevant optimization and implementation challenges to achieve this goal are also addressed.

The rest of this chapter is organised as follows. In the introductory Section 4.1, we outline the goals to be achieved and the contributions presented in this chapter. Section 4.2 presents our preliminary anonymization scheme which was improved in Section 4.3. In Section 4.4 the implementation issues are discussed that are validated with simulation in Section 4.5. Finally, Section 4.6 concludes the chapter.

4.1 Introduction

The combined challenges of protecting location privacy of the participants in a PSS and, at the same time, ensuring a high quality of data at the desired end are significantly difficult.

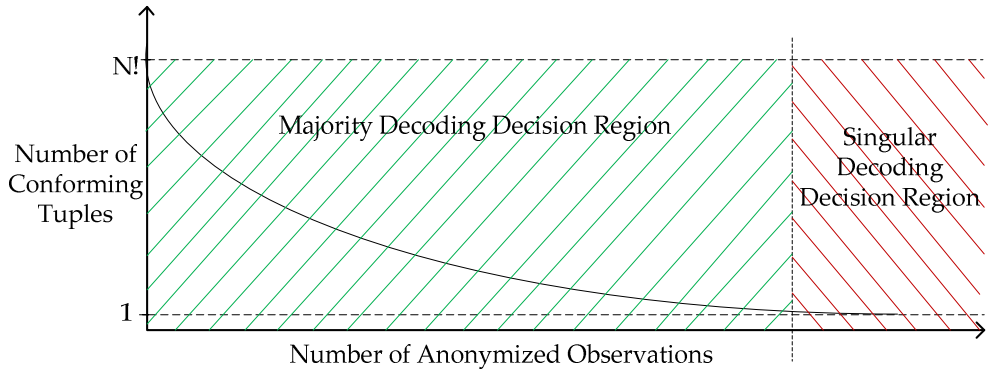


Figure 4.1: Two different working regions of proposed greedy techniques.

Most of the existing techniques try to address this challenges by designing anonymization algorithms to determine spatial cloaks containing at least k POIs. For example, in [36] a user reports the centre of a tile consisting of at least k POIs or alternatively their mean location as the location of the reported POI. The server associates this report with the nearest POI in the tile. Clearly, this work did not consider the potential damage of data quality. It has been shown in [34] that the proposed solution in [36] can achieve on average $1/k$ data integrity, i.e., the server is able to correctly associate POIs on average $1/k$-th time, which is a significantly poor performance especially for reasonable high anonymity. In our case, the subset-coding technique is used so that each observation from a participant can be transmitted with sufficient anonymity, whereas the data collector can de-anonymize individual data only through the joint decoding of the entire collection.

Since some of the POIs in many PSS application scenarios may be located in remote places, the number of observers of those is likely to be small. Furthermore, the observed attribute is likely to vary at certain intervals. For example, in *PetrolWatch*, the price of petrol fluctuates quite frequently. These add more complexity to the problem that the challenges need to be satisfied with a small number of observations. Consequently, the anonymization algorithm has to address new dimensions while forming each individual AR. The AS would perform this task by using the knowledge of already generated ARs. In this chapter, we present such an anonymization algorithm using a probabilistic greedy heuristic that aims to provide sufficient data quality at the ApS even with a small number of ORs. Note that all the terms AR, AS, ApS, OR were introduced and defined in Chapter 3 and retain the same meaning throughout this thesis.

From the basic concept of joint decoding, as illustrated in the previous chapter, we know that the more anonymized observations are received the number of conforming tuples are

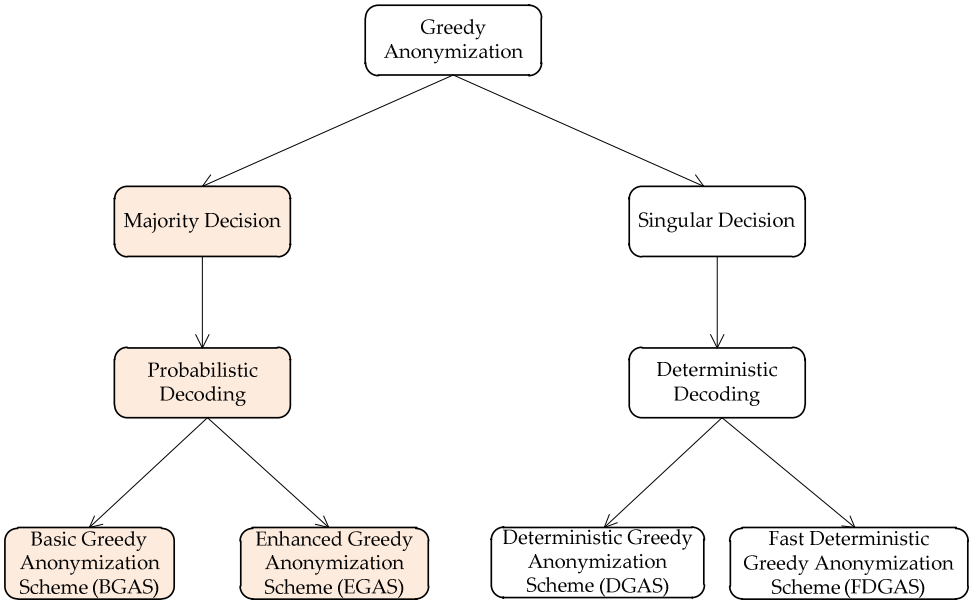


Figure 4.2: Two different approaches of proposed greedy techniques.

reduced. Using our greedy heuristics we aim to optimize the anonymization techniques such that the joint subset decoding performance can be maximized with the minimum number of observations. Figure 4.1 shows how the decision making on the basis of joint subset decoding is dictated by the current status of total number of conforming tuples. Before reaching a single line we can make a decision probabilistically, which aims to reflect the actual scenario for the majority of time during the service. When it reaches single conformity, we can make a singular decision deterministically. We want to optimize the majority decision probabilistically and the singular decision deterministically (as shown in Figure 4.2). This chapter will focus on the probabilistic approach and the deterministic one will be discussed along with a comparative analysis in the next chapter.

The preliminary concept of subset coding and its application by greedy anonymization was first introduced by us in [33]. We titled this probabilistic approach *Basic Greedy Anonymization Scheme* (BGAS). At the time, the necessity for enhancement of this scheme was gained from our observation that to ensure the highest degree of anonymity, i.e., $k = N - 1$, the number of observations required to decode the correct POI-attribute association may be deemed impractical. From this consideration, we explored several optimization issues and developed an *Enhanced Greedy Anonymization Scheme* (EGAS) [34]. Both the schemes use probabilistic methods to perform the decoding and, hence, are jointly named the *Probabilistic Greedy Anonymization Scheme* (PGAS). These use a majority decoding

technique which operates based on the ability of matching the correct attribute with the POI in majority of the cases. In this chapter, we present both these schemes and compare their performance in terms of their achieved data quality. Our key contributions in this chapter are outlined below:

- Developing a probabilistic greedy anonymization algorithm BGAS to achieve high data quality after joint-decoding at the target end,
- Improving the anonymization to EGAS by designing a number of optimization strategies to ensure data quality even with a small number of observations,
- Analysing the transient impact of change in the attribute of POIs.

Finally, extensive simulation results are presented to establish the applicability of our proposed approach.

Now, we present the proposed greedy anonymization techniques and also show how the anonymized reports are decoded jointly, using the concepts of subset-coding and joint decoding presented in Section 3.3.

4.2 BGAS

In this section, we first present the concept of BGAS. We then present BGAS formally along with the algorithms for its anonymization scheme and decoding scheme. Section 4.2.1 provides a brief introduction to the concept of BGAS followed by relevant algorithms in Section 4.2.2. Finally, in Section 4.2.3 our proposed scheme is compared to a contemporary approach from the viewpoint of data integrity performance.

4.2.1 Concept of BGAS

For anonymization of each report from MN, the AS will generate a new subset. This tries to augment the already developed optimal subsets derived from past reports, provided that in using the subsets developed so far, the actual price retrieval performance can be maximized overall. Let there be N numbers of POIs whose prices are uniquely defined. Our AS receives the actual price and the corresponding POI from the MN. To make a report with k-anonymity, the actual POI can be anonymized with any $k-1$ out of the remaining $N-1$ POIs. For the very first input of price, the AS has nothing much to do and it can randomly pick any combination to construct the AR. From the second input and onwards, it is found that carefully picking $k-1$ POIs can lead to significant revelation towards the actual prices

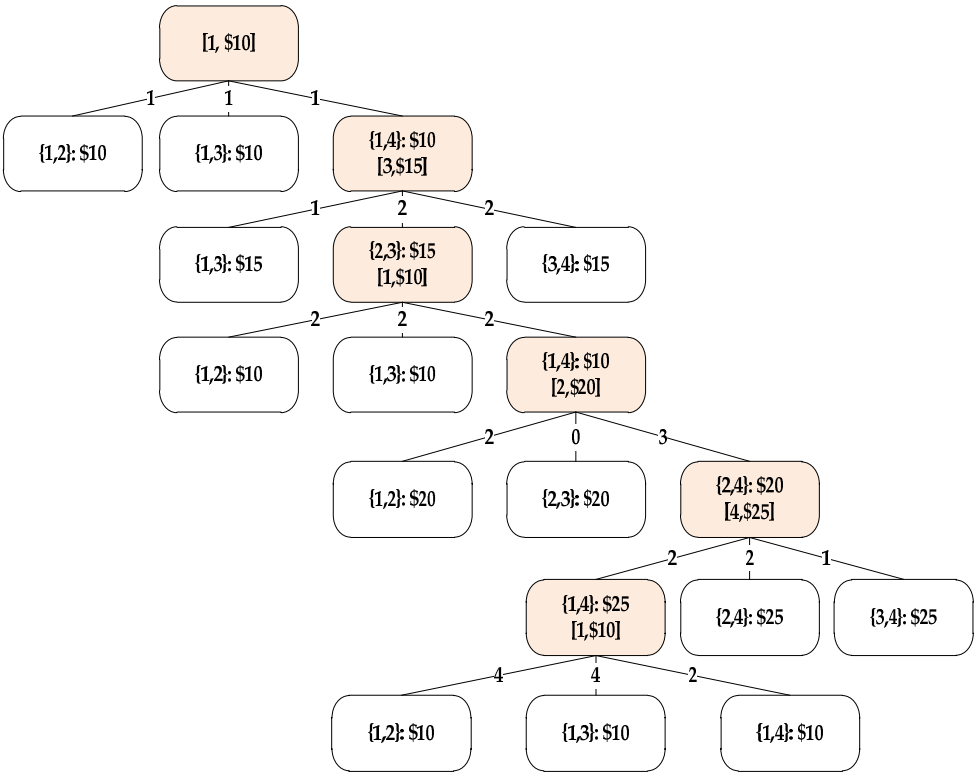


Figure 4.3: Subset generation procedure in BGAS.

of the POIs. Our algorithm takes the advantage of this hypothesis. From all the possible subsets, it selects the subset which, after being added to the subsets already sent to the ApS, gives the maximum match of deduced prices against the actual ones. In this way, individual location privacy is preserved, while at the end the application server can also declare the prices of the POIs with an acceptable level of data integrity.

Considering the example of *PetrolWatch* with $N = 4$ and $k = 2$, let us observe Figure 4.3 where one-by-one six ARs are generated in the AS with six incoming ORs. In the figure, (i) the root represents the first observation; (ii) each node in a level represents one of the possible ARs for the observation made in the above level, and (iii) the next observation is written along with the selected AR in that level. The solid edges of the tree refer to the number of match of prices for the given OR. It is clear that at each step the AR with the maximum match is selected to be appended with the next level subset derived from the next input. In the case of more than one subset exhibiting a maximum match, any one is chosen to be the selected subset.

Let us assume that the AS has already generated five ARs using this approach. For a better understanding we have taken a very simple example. Now we will see how the algorithm actually works by observing the selection and introduction of the sixth AR upon receiving the sixth OR. Let the first five OR be as follows:

$$[1, \$10] \quad [3, \$15] \quad [1, \$10] \quad [2, \$20] \quad [4, \$25]$$

The anonymized and selected ARs so far reported for these five ORs are:

$$\{1,4\}: \$10 \quad \{2,3\}: \$15 \quad \{1,4\}: \$10 \quad \{2,4\}: \$20 \quad \{1,4\}: \$25$$

Only the following price permutations satisfy the conditions in all the above five ARs:

POI 1	*POI* 2	*POI* 3	*POI* 4
\$10	\$20	\$15	\$25
\$25	\$20	\$15	\$10

Then comes the sixth OR as $[1, \$10]$. For achieving 2-anonymity the AS can consider anyone of the following three possible ARs:

$$\{1,4\}: \$10 \quad \{1,2\}: \$10 \quad \{1,3\}: \$10$$

The task of our algorithm is to check which AR leads to the maximal actual price retrieval ratio after being added to the five ARs and then to select that one.

At first the AR $\{1,4\}: \$10$ is considered. It will only pick the permutations that validate either $[4, \$10]$ or $[1, \$10]$. Thus, the table remains the same as before and the number of price occurrences against the POIs is given in the following association matrix. According to this matrix, the prices of POIs 1 and 4 are indeterminable, but the prices of POIs 2 and 3 correspond to the actual, leading to two matches.

	POI 1	*POI* 2	*POI* 3	*POI* 4
\$10	1	0	0	1
\$20	0	2	0	0
\$15	0	0	2	0
\$25	1	0	0	1

After applying the second possible AR $\{1,2\}: \$10$ or the third one $\{1,3\}: \$10$ only the following permutation survives:

POI 1	*POI* 2	*POI* 3	*POI* 4
\$10	\$20	\$15	\$25

With the following association matrix:

	POI 1	*POI* 2	*POI* 3	*POI* 4
\$10	1	0	0	0
\$20	0	1	0	0
\$15	0	0	1	0
\$25	0	0	0	1

Clearly, the prices of all POIs can be decoded, leading to four matches.

From the above results, it is clearly evident that both the ARs {1,2}: \$10 and {1,3}: \$10 can give the best revelation towards the actual prices of the POIs. Therefore our greedy algorithm selects the best performing AR considered first, that is, {1,2}: \$10 in this step and follows the same approach for the next reports.

4.2.2 The Decoding and Anonymization Algorithms

Now, we formally present BGAS. The approach is a greedy one as it picks the best choice of the next rule in every step. The scheme basically comprises two algorithms: one for anonymized rule generation and the other for decoding these anonymized messages. On the one hand, the decoding algorithm, Algorithm 4.1 will be used by ApS and also by the adversaries. On the other hand, AS will use the anonymization algorithm, Algorithm 4.2, which inherently uses the decoding algorithm. Before describing the algorithms in detail, we first define some essential terms.

Definition 4.1 (Possible Anonymization Subsets): *Possible Anonymization Subsets (PAS) are the anonymized subset parts of all possible ARs corresponding to a particular POI for a fixed k and N. The total number of PAS corresponding to a POI is $\binom{N-1}{k-1}$.*

$$PAS_i^{\{1,\dots,N\},k} = \begin{cases} \emptyset, & \text{if } k \geq |N| \vee i \notin \{1,\dots,N\} \\ \{\{i, i_1, \dots, i_{k-1}\} | \{i_1, \dots, i_{k-1}\} \subset \{1,\dots,N\} \setminus \{i\}\}, & \text{otherwise} \end{cases} \quad (4.1)$$

Definition 4.2 (Possible Attribute Assignment Set): *A Possible Attribute Assignment Set (PAAS) is a set of all possible permutations of the known attributes with corresponding POIs. Each individual of this set is a possible POI-attribute association permutation. So, the mathematical notation of PAAS can be denoted as,*

$$P \equiv \langle p_1, p_2, \dots, p_N \rangle, \quad (4.2)$$

where $p_i \in \{a_1, a_2, \dots, a_N\} \wedge p_i \neq p_j$, iff $i \neq j$.

Definition 4.3 (Conforming Attribute Assignment Set): *A Conforming Attribute Assignment Set (CAAS) is a subset of PAAS where conforming N price tuples are listed eliminating the non-conforming ones, each time an AR is generated. So, $CAAS_{AR} \subset PAAS$ and the operation of checking the conformity of a generated AR is performed as follows,*

$$P \oplus AR_i = \begin{cases} P, \ if \bigvee_{j=1}^{k} (p_{i_j} = a_i) ; \\ \emptyset, \ \text{otherwise.} \end{cases} \quad (4.3)$$

It means that when found conforming to the generated AR, the set of permutation is returned as it is. Otherwise, this operation returns a null set. Each permutation of PAAS is checked against a generated AR and only the conforming ones constitute the CAAS as given below,

$$CAAS_{AR_i} = \bigcup_{j=1}^{N!} PAAS_j \oplus AR_i \quad (4.4)$$

Definition 4.4 (Anonymized Rules Set): *An Anonymized Rules Set (ARS) is a set of generated*

Algorithm 4.1: $(da_1, \dots, da_N) =$ **Decode_BGAS** (N, k, ARS)

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Set of ARs, $ARS = (AR_1, \dots, AR_m)$ where $AR_j \equiv \{i_{j1}, \dots, i_{jk}\}: a_{i_j} | i_j \in \{i_{j1}, \dots, i_{jk}\}$ for all $1 \leq j \leq M$

Output:

- Decoded attributes, $(da_1, \dots, da_N)$.

1. Set $A = \{\alpha | \exists ss \in PAS: (ss: \alpha) \in ARS\}$
2. IF $|A| < N$ THEN
3. Set $A = A \cup \{-1, \dots, -(N - |A|)\}$
4. END IF
5. Let $(\alpha_1, \dots, \alpha_N)$ be any arbitrary ordering of A
6. Set $C = \{c_1, \dots, c_{N!}\}$, where $c_i = (c_{i,1}, \dots, c_{i,N})$ is a unique permutation of $(\alpha_1, \dots, \alpha_N)$ for all i
7. FOR each $AR \in ARS$ DO
8. Set $C = C \oplus AR$
9. END FOR
10. IF $|C| > 0$
11. FOR $i = 1, \dots, N$ DO
12. Set $da_i = \begin{cases} \text{mode}(\{c_{1,i}, \dots, c_{|C|,i}\}), & \text{if mode}(\{c_{1,i}, \dots, c_{|C|,i}\}) \geq 0 \text{ and unique} \\ -\infty, & \text{otherwise} \end{cases}$
13. END FOR
14. ELSE
15. Set $da_i = \infty$ for all $1 \leq i \leq N$
16. END IF

ARs selected so far by the greedy algorithm of AS.

$$CAAS_{ARS} = CAAS_{\{AR_1,\dots,AR_j\}} = \bigcap_{i=1}^{j} CAAS_{AR_j}. \tag{4.5}$$

Each time a new AR is generated and is included in the ARS, the CAAS is updated as given below,

$$CAAS_{ARS\cup\{AR\}} = CAAS_{ARS} \cap CAAS_{AR}. \tag{4.6}$$

Algorithm 4.1 collects all attributes and adds dummies (negative values), if necessary (steps 1-5). It then disregards the permutations that are non-conforming to any $AR \in ARS$ from the set of all possible permutations (steps 6-9). Finally, it decodes for maximal probability using unique majority or indicate not-yet-decodable (steps 10-12) or declare contradiction (steps 15) when there no permutation survives.

Lemma 4.5. *Computational complexity of Algorithm* 4.1 *is* $O(N^N)$.

Algorithm 4.2: AR = **Anonymize_BGAS** $(N, k, \&(\alpha_1, \dots, \alpha_N), \&ARS, o \equiv [i, a_i])$

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Actual attributes of all POIs from recent observations, $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, α_i is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values
- Set of already generated ARs, ARS
- Observation report $o \equiv [i, a_i]$, $1 \le i \le N$

Output:

- Anonymized rule $AR \equiv \{i_1, \dots, i_k\}: a_i | i \in \{i_1, \dots, i_k\}$

1. IF $\alpha_i \ge 0$
2. Set $a_i = -a_i$
3. Set $ARS = \emptyset$
4. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
5. END IF
6. Set $\alpha_i = |a_i|$
7. Set $ss = \underset{\forall ss \in PAS_i^{\{1,\dots,N\},k}}{\operatorname{argmax}} |\{\alpha_t | 1 \le t \le N \wedge \alpha_t \ge 0 \wedge da_t = \alpha_t\}|$
   where $(da_1, \dots, da_N) = \mathbf{Decode\_BGAS}(N, k, ARS \cup \{(ss: \alpha_i)\})$
8. Set $AR = (ss: a_i)$
9. Set $ARS = ARS \cup \{AR\}$

Proof. : Let the time complexity of the algorithm be $T(N)$. Lines 11 – 13 represent the most dominating computation. Line 12 takes $O(N!)$ time as $|C| = N!$. This operation is repeated for N times as shown in line 11. Hence, the total complexity of the algorithm, $T(N) = O(N \times N!) = O(N^N)$. ■

The anonymization algorithm starts with a null set of already generated ARs, ARS. Then, the algorithm 4.2 removes all past observations when any attribute fluctuation is detected, which is then reported by encoding the attribute with a negative sign (steps 1-5). Then, it records the new attribute in step 6. Finally, it constructs AR using a subset that achieves maximum-possible decodability using maximal probability decoding (steps 7-9).

Lemma 4.6. *Computational complexity of Algorithm* 4.2 *is* $O(N^N)$.

Proof. : Let the time complexity of the algorithm be $T(N)$. Line 7 represents the most dominating computation. Here, the decoder is called for each subset in PAS and $\left|PAS_i^{\{1,\dots,N\},k}\right| \leq \binom{N-1}{k-1} = O(2^N)$, for $k = \frac{N}{2}$ as it represents the worst-case scenario. From Lemma 4.5, the time complexity of **Decode_BGAS**(N, k, ARS) is $O(N^N)$ Hence, the total complexity of the algorithm, $T(N) = O(2^N \times N^N) = O(N^N)$. ■

As we have discussed earlier, AS covers a small locality with a small number of particular types of POIs like petrol station, hospital, super store, etc. are natural. Therefore, the value of N should not exceed 7 and hence it may handle this high computational complexity.

4.2.3 Data Integrity Performance

Among existing privacy preservation techniques, we chose the one proposed by Huang *et al.* [36] to compare with as their proposed scheme also applied *k*-anonymization for protecting location privacy. A brief description of their method was given in Chapter 2 and in that scheme k POIs were reported by a single point (centre of their tile or alternatively the mean of the equivalence classes). Consequently, only $1/k$ POI was correctly reported, which is the actual price retrieval rate of that scheme. The detail of the performance comparison is presented in Section 4.5.2.

4.3 EGAS

In this section, we first present the concept of EGAS. Then we investigate a number of optimization issues to improve the quality of decoded data and present EGAS formally along with a discussion on how it differs from BGAS. Section 4.3.1 gives a brief introduction to the concept of EGAS. Some optimization issues are addressed in Section 4.3.2 followed by relevant algorithms in Section 4.3.34.3.3.

4.3.1 Concept of EGAS

BGAS and EGAS differ in their methods of AR selection. In BGAS, ARs are selected such that maximal POIs can be associated by majority consideration with their correct attributes. In contrast, EGAS select ARs such that the currently observed POI can be associated with its actual attribute by the majority. First, we need to find all possible N-tuples of attributes conforming to all ARs in the set. Then, each POI is associated with the majority attribute in those tuples. If there is more than one majority attribute, the POI is associated with none. Consider the same example of *PetrolWatch* with four POIs having prices of \$10, \$20, \$15, and \$25, respectively, as considered for BGAS in Section 4.2.1. Let us assume that the same

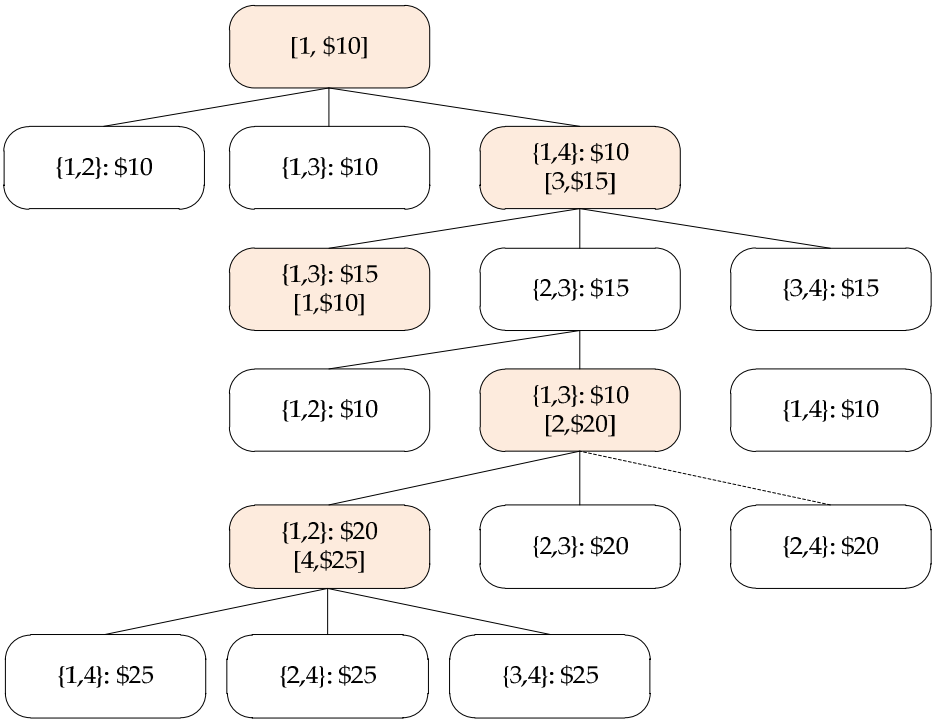


Figure 4.4: Subset generation procedure in EGAS.

observation sequence is considered here as was considered when illustrating BGAS, which is, $[1, \$10]$, $[3, \$15]$, $[1, \$10]$, $[2, \$20]$, $[4, \$25]$, and $[1, \$10]$. In response to this how the AS generated corresponding 2-anonymous ARs is demonstrated here. Note that the term decodability is defined in Definition 3.3.

Figure 4.4 depicts a concise representation of the above example in the form of a tree where (i) the root represents the first observation; (ii) each node in a level represents one of the possible ARs for the observation made in the above level; and (iii) the next observation is written along with the selected AR identified by the darker node in that level. Whether an observation is correctly decoded or not using the set of ARs following the path from the root up to a node in the level below the observation is denoted respectively by a solid or dashed link to that node. If a node is repeated in a level above to it, no link is drawn.

As compared to the six ARs required in BGAS to reach full decodability, here only five 2-anonymous ARs $\{1,4\}: \$10$, $\{1,3\}: \$15$, $\{1,3\}: \$10$, $\{1,2\}: \$20$, and $\{1,4\}: \$25$ for the respective ORs are found enough to reach full decodability, i.e., all four POIs can be associated with the correct price.

Let us now demonstrate how a set of ARs for price reports can be generated by EGAS so that 4-decodability is achieved. When the first observation report $[1, \$10]$ is received by the AS, it uses three dummy prices so that the possible ARs can be ranked based on whether or not the observed POI can be associated with the correct price. As all the three possible ARs,$\{1,2\}: \$10$, $\{1,3\}: \$10$, and $\{1,4\}: \$10$ can correctly associate price \$10 to POI 1 using majority decoding (in each case, there are 12 conforming 4-tuples with the majority of six having \$10 associated with POI 1), the AS randomly selects the AR $\{1,4\}: \$10$ in response to this OR.

When the second OR $[3, \$15]$ arrives, the AS needs to use only two dummy prices. Again, all three possible ARs,$\{1,3\}: \$15$, $\{2,3\}: \$15$, and $\{3,4\}: \$15$ can correctly associate the price \$15 to POI 3 using majority decoding when used in conjunction with the previously generated AR $\{1,4\}: \$10$. Hence, the AS randomly selects the AR $\{1,3\}: \$15$ in response to this observed report.

As the third OR $[1, \$10]$ is a repeat of the first received OR, the AS still needs to use the two dummy prices. This time, out of the three possible ARs, the previously generated AR $\{1,4\}: \$10$ is excluded from the choice. This is shown in Figure 4.4 by drawing no link to the

Table 4.1: Conforming Tuples of Gradually Generated AR Set where Dummy Attributes are Identified with Leading d

ARS	{1,4}: $10	{1,4}: $10 {1,3}: $15	{1,4}: $10 {1,3}: $15 {1,3}: $10	{1,4}: $10 {1,3}: $15 {1,3}: $10 {1,2}: $20	{1,4}: $10 {1,3}: $15 {1,3}: $10 {1,2}: $20 {1,4}: $25
$CAAS_{ARS}$	$(\$10, d1, d2, d3)$ $(\$10, d1, d3, d2)$ $(\$10, d2, d1, d3)$ $(\$10, d2, d3, d1)$ $(\$10, d3, d1, d2)$ $(\$10, d3, d2, d1)$ $(d1, d2, d3, \$10)$ $(d1, d3, d2, \$10)$ $(d2, d1, d3, \$10)$ $(d2, d3, d1, \$10)$ $(d3, d1, d2, \$10)$ $(d3, d2, d1, \$10)$	$(\$10, d1, \$15, d2)$ $(\$10, d2, \$15, d1)$ $(d1, d2, \$15, \$10)$ $(\$15, d1, d2, \$10)$ $(\$15, d2, d1, \$10)$ $(d2, d1, , \$15, \$10)$	$(\$10, d1, \$15, d2)$ $(\$10, d2, \$15, d1)$	$(\$10, 20, \$15, d1)$	$(\$10, 20, \$15, \$25)$

AR excluded from consideration. As the remaining two possible ARs can correctly associate price \$10 to POI 1 by majority, the AS randomly picks up AR $\{1,3\}: \$10$.

When the fourth observation $[2, \$20]$ arrives the AS now knows three prices and hence needs to use only one dummy price. Among the three possible ARs only the two ARs $\{1,2\}: \$20$ and $\{2,3\}: \$20$ can correctly associate the price \$20 to POI 2 when used together with the previously generated three ARs. The third AR is shown in Figure 4.4 by drawing a dashed link to the AR not fulfilling the majority decoding requirement. Let us assume that the AS randomly picks the AR $\{1,2\}: \$20$ in response to this observation.

When the final observation $[4, \$25]$ arrives, AS no longer needs any dummy price and any of the three possible ARs can associate the price \$25 to POI 4 correctly by the majority decoding when is used along with the previously generated four ARs. Indeed, these five ARs can now successfully associate all POIs to correct prices.

In Table 4.1, conforming price 4-tuples are shown each time an AR is generated. Note that the number of tuples gradually decreases with the number of ARs. By adding a new AR, some of the existing conforming tuples no longer remain conforming to the new set of

ARs. Effectively, only the tuple representing correct POI-attribute association for all POIs survives.

4.3.2 Optimization Issues in EGAS

In this section, we investigate three aspects of optimization to improve decodability of the data at ApS with fewer observations. Section 4.3.2.1 presents a theorem with its proof followed by some reasons behind choosing the local search direction in Section 4.3.2.2. Finally, in Section 4.3.2.3, another intuitive optimization issue of joint anonymization of multiple observations is explored.

4.3.2.1 Recurrent Observations

In BGAS, all possible $\binom{N-1}{k-1}$ ARs are considered for each new observation, whereas EGAS only considers ARs that are not included in the generated set of ARs. This is significant in reducing the total number of observations needed to achieve full decodability with a high degree of anonymity. The right-most node in level 3 of the AR generation tree in Figure 4.4 is not considered by EGAS despite being capable of associating price \$10 with the recently observed POI 1. BGAS will pick up this node with a probability of $1/3$. With a large N and k, there will be many such repeating nodes in the tree and the probability of picking these nodes by BGAS increases accordingly. Selecting repeating ARs by the AS cannot improve decodability as proved by the following theorem.

Theorem 4.7. *Performance of the decoder is independent of the order of ARs and duplicate ARs do not improve the performance.*

Proof. Irrespective of the order of ARs, the conforming attribute N-tuples are fixed for the set of ARs and duplicate ARs neither make any non-conforming tuple conforming nor does it cause any conforming tuple to become non-conforming. ■

Note that in our example, had AR $\{1,4\}$: \$10 been selected for both the observations of POI 1, the set of conforming tuples would have remained unchanged when the second observation is processed, whereas the proposed EGAS successfully reduced the size of the set from 4 to 3 as the 4-tuple $(d1, \$20, \$30, \$10)$ no longer remains conforming as shown in Table 4.1.

4.3.2.2 Local Search Direction

In selecting the best AR from a list of possible ARs, BGAS selects the one that would lead to associate the maximum number of POIs with their correct attributes by majority decoding, whereas EGAS selects the one that would guarantee the correct association of attributes to the currently observed POI only. From a different viewpoint, the greedy local search in BGAS and EGAS is performed respectively with global and local optima objectives.

A greedy algorithm is a problem solving heuristic of making the locally optimal choice at each stage of a process with the hope of finding the global optimum. Although the greedy algorithm often fails to produce the optimal solution when it reaches a local plateau, this is not the case if the problem space has a convex surface such as the problem of subset anonymization. In this problem, the number of conforming tuples is monotonically non-increasing irrespective of the search technique used. A hill-climbing greedy algorithm such as EGAS, which attempts to find a better solution by incrementally changing a single element of the solution, is well-suited for optimizing over convex surfaces and converges to the global maximum quicker [37]. Whether performing a local search with a global or local optima objective makes any substantial difference in the final outcome depends on the problem in hand and any analytical modelling to justify the preference is hard. Instead we opt for the following empirical analysis.

As observations are made at random with independent observers, the probability distribution of a sequence of M observed POIs is difficult to model. However, we can use the following two specific sequences to represent the ideal and extreme cases:

$$Seq_{ideal}(M) = \langle \overbrace{1,2,\dots,N,1,2,\dots}^{M} \rangle. \tag{4.7}$$

$$Seq_{extreme}(M) = \langle \overbrace{1,\dots,1}^{\lceil M/N \rceil}, \overbrace{2,\dots,2}^{\lceil M/N \rceil}, \dots, \overbrace{N,\dots,N}^{\lfloor M/N \rfloor} \rangle. \tag{4.8}$$

The proposed algorithm selects an AR at random from a set of possible ARs that maximises the selection criteria (global or local optima objective). As a result, a sequence of observation can result in many possible sets of ARs depending on the selection criteria. It is observed that using global optima objective prunes the search tree significantly at the beginning, whereas using local optima objective prunes the search tree at a much lower level. Consequently, the number of possible sets of ARs for a given sequence of observation is much smaller when local optima objective is used.

Table 4.2: Mean Decodability and Percentage of Iterations Achieving N-Decodability for the Ideal and Extreme Sequences of Observed POIs with $N = 4$ and $k = 3$

Sequence of Observed POIs	Mean Decodability		Percentage of Full Decodability	
	Local	Global	Local	Global
Ideal: ⟨1,2,3,4,1,2,3,4,1⟩	3.103	2.167	58.7	8.3
Extreme: ⟨1,1,1,2,2,3,3,4,4⟩	3.789	3.382	90.3	69.1

We have generated ARs for the above two sequences of various lengths with different values of N and k. In all cases, tens of thousands of iterations were carried out and the mean decodability and the percentage of iterations achieving full decodability were estimated. Although we observed the same trend in all cases, for the sake of brevity, we present the results for $N = 4$, $k = 3$, and $M = 9$ in Table 4.2. Clearly, using local optima objective, as we have introduced in EGAS in which the correct association of attribute to only the currently observed POI is checked, is superior to using global optima objective where the maximum number of POI-attribute association is preferred. Note that the performance of the local objective is significantly better in the ideal sequence of observations. In Section 4.5.3, we will present a detailed comparative performance analysis on these two search directions.

4.3.2.3 Joint Anonymization of Multiple Observations

While demonstrating the basic concept in Section 4.3.1, we have anonymized observation reports one at a time. A natural optimization enquiry is to find whether decodability performance can be significantly improved if λ successive observations are grouped and jointly anonymized. This can be performed by exploring λ levels of the AR generation tree to select λ ARs that satisfy the local optimal criteria, which needs to be modified to find the maximum of λ POIs associated to correct attributes.

Table 4.3: Mean Decodability and Percentage of Iterations Achieving N-Decodability with Joint Anonymization for the Ideal and Extreme Sequences of Observed POIs with $N = 4$ and $k = 3$

Sequence of Observed POIs	Mean Decodability			Percentage of Full Decodability		
	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Ideal: ⟨1,2,3,4,1,2,3,4,1⟩	3.103	3.484	3.742	58.7	74.2	87.1
Extreme: ⟨1,1,1,2,2,3,3,4,4⟩	3.789	3.233	2.166	90.3	66.0	14.3

We have performed simulations with the same set up used in the previous section. Again, tens of thousands of iterations were carried out and the mean decodability and the percentage of iterations achieving full decodability were estimated. Although we observed the same trend in all cases, for the sake of brevity, we present the results for $N = 4, k = 3, M = 9$, and $\lambda \in \{1,2,3\}$ in Table 4.3. Higher values of λ were avoided due to a significant increase in decoding complexity. This time, however, we were unable to observe any superior choice. While decodability improves with λ for the ideal sequence of observation, the opposite trend was observed for the extreme sequence. This can be explained as follows. No doubt that higher λ ensures improved AR set for any independently selected λ observations. However, observations used by our greedy scheme cannot be considered

Algorithm 4.3: AR = **Anonymize_EGAS** $(N, k, \&(\alpha_1, \dots, \alpha_N), \&ARS, o \equiv [i, a_i])$

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Actual attributes of all POIs from recent observations, $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, α_i is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values
- Set of already generated ARs, ARS
- Observation report $o \equiv [i, a_i]$, $1 \leq i \leq N$

Output:

- Anonymized rule $AR \equiv \{i_1, \dots, i_k\}: a_i | i \in \{i_1, \dots, i_k\}$

1. IF $\alpha_i \geq 0$
2. Set $a_i = -a_i$
3. Set $ARS = \emptyset$
4. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
5. END IF
6. Set $\alpha_i = |a_i|$
7. Set $SS = \left\{ ss \,\middle|\, \begin{matrix} ss \in PAS_i^{\{1,\dots,N\},k} \wedge da_i = \alpha_i \\ \text{where } (da_1, \dots, da_N) = \textbf{Decode\_BGAS}(N, k, ARS \cup \{(ss: \alpha_i)\}) \end{matrix} \right\}$
8. IF $SS \neq \emptyset$ THEN
9. IF $|SS| = 1$ THEN
10. Set $AR = (ss: a_i)$ where $ss \in SS$
11. ELSE
12. Set $AR = (ss: a_i)$ where $ss \in SS \wedge (ss: \alpha_i) \notin ARS$
13. END IF
14. ELSE
15. Set $AR = (ss: a_i)$ where $ss \in PAS_i^{N,k}$
16. END IF
17. Set $ARS = ARS \cup \{AR\}$

independent as we accumulate ARs as observations are being reported. Consequently, there can no longer be any guarantee of improved decodability with $\lambda > 1$ as the set of the already generated ARs no longer remains the same when $\lambda = 1$.

Nevertheless, computational complexity of the anonymization algorithm increases significantly with λ. Moreover, the decodability performance for the extreme sequence degrades at a significantly higher rate as λ is increased. Hence, $\lambda = 1$ is preferred. For the sake of completeness, we will a present detailed comparative performance analysis on joint anonymization in Section 4.5.4.

4.3.3 The Anonymization and Decoding Algorithms

To anonymize with the aim of achieving high data integrity, it is essential to inherently

Algorithm 4.4: $(da_1, \dots, da_N) = \textbf{Decode\_EGAS}\big(N, k, \&(\alpha_1, \dots, \alpha_N), \&C, AR \equiv (ss: a)\big)$

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Actual attributes of all POIs from recent observations, $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, α_i is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values.
- Updated $CAAS_{ARS}$, C
- Anonymized Rule, $AR \equiv (ss: a)$

Output:

- Decoded attributes, $(da_1, \dots, da_N)$

1. IF $C = \emptyset \vee a < 0 \vee \big(\forall i: \alpha_i \geq 0 \wedge \nexists j: \alpha_j = a\big)$ THEN
2. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
3. Set $C = \{c_1, \dots, c_{N!}\}$ where $c_i = \big(c_{i,1}, \dots, c_{i,N}\big)$ is a unique permutation of $(\alpha_1, \dots, \alpha_N)$ for all i
4. END IF
5. Set $\alpha_j = |a|$ where $j = \underset{\forall i}{\operatorname{argmin}}\ \alpha_i < 0$
6. Set $C = C \oplus AR$
7. IF $C \neq \emptyset$ THEN
8. FOR $i = 1, \dots, N$ DO
9. Set $da_i = \begin{cases} \text{mode}\big(\{c_{1,i}, \dots, c_{|C|,i}\}\big), & \text{if mode}\big(\{c_{1,i}, \dots, c_{|C|,i}\}\big) \geq 0 \text{ and unique} \\ -\infty, & \text{otherwise} \end{cases}$
10. END FOR
11. ELSE
12. Set $da_i = \infty$ for all $1 \leq i \leq N$
13. END FOR

decode tentative rules. This is done to check the decodability performance of each candidate AR upon which the decision of final selection is made. Hence, the decoder called from the AS should be offline. However, we prefer to use a real-time decoder for decoding at the ApS end such that any attempt of false data feeding or actual attribute fluctuations are readily detected. Here, we first present the anonymization scheme (Algorithm 4.3) which inherently checks decoding performance of the tentative ARs using the offline decoder (Algorithm 4.1) and then the online decoding scheme (Algorithm 4.4) used in ApS.

Algorithm 4.3 receives an observation report and the set of actual attributes accumulated in AS. Note that steps 1 to 6 are just the same as in algorithm 4.2. Both these anonymization algorithms differ in the step of selecting the AR. However, another significant difference in step 7 is that the ARs that were generated earlier for the same observation are excluded from this set of PAS here. Finally, the greedy algorithm constructs AR using a subset, preferably non-repeated, that can decode the current attribute using maximal probability decoding, if possible (steps 7-17). Its worst-case computational complexity is the same as that of algorithm 4.2 i.e., $O(N^N)$ as proved in Lemma 4.6. However, we can design the system in such a way that every locality will contain multiple ASs, each having a small number of POIs to deal with. Still, in scenarios where this system design is not feasible, we may use more efficient ones. These are discussed in the next chapter.

Algorithm 4.4 receives a set of ARs and finds the set of attributes from these ARs. The decoder is reset at the beginning or when the previous decoding contradicts due to a malicious AR or the AS signals attribute fluctuation or number of attributes exceeds N due to false data feeding (steps 1-4). Then, it removes permutations non-conforming to AR after recording the new attribute (steps 5, 6). Finally, it decodes for maximal probability using the unique majority or indicates not-yet-decodable (steps 7-10) or declare contradiction when no permutation survives (step 12).

The computational complexity of this online decoding scheme is the same as that of algorithm 4.1, i.e., $O(N^N)$ as proved by Lemma 4.5. However, this computational complexity analysis is considered the worst case scenario. For the average case scenario, we may analyse the complexity as follows. For each unique AR, the decoder scans the list of so-far-conforming tuples to eliminate the ones that are non-conforming to AR. The average-case order of the decoder will be expressed in terms of the average number of permutations scanned for each AR.

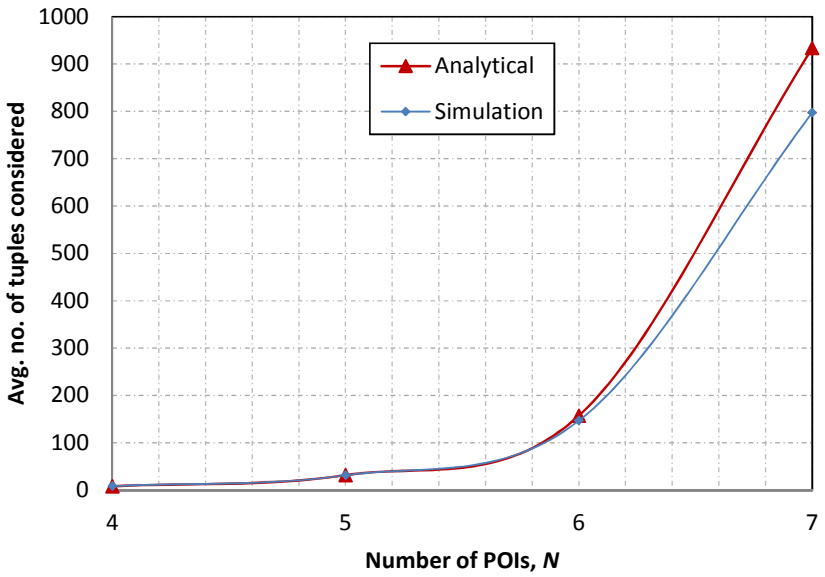


Figure 4.5: Average-case decoding complexity in EGAS.

The decoder starts at state $\mathbb{S} = N$ undecided attributes, with $N!$ permutations. For the sake of simplicity, we assumed that the decoder groups ARs involving the same attribute and considers the groups in order of AR frequency. We may then generalise that the most-observed attribute will be decoded using $a_{\mathbb{S}=N} = \max(N-1,1)$ ARs by reducing the number of conforming permutations to $(N-1)!$ and hence, transiting the decoder to the next state $\mathbb{S} = N-1$ undecided attributes. By further assuming that each of these ARs eliminates the same $\frac{N!-(N-1)!}{a_{\mathbb{S}=N}} = (N-1)!$ number of non-conforming tuples, we may once again generalise that each of these ARs scans $N! - (j-1)(N-1)!$ permutations where $1 \leq j \leq a_{\mathbb{S}=N}$ is the rank of the AR in the observational temporal order.

Considering that the decoder will gradually transit from the initial state $\mathbb{S} = N$ to the final state $\mathbb{S} = 1$ to achieve full-decodability, we can estimate the average number of tuples scanned for each AR to be

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{a_i-1}(i! - (j-1)(i-1)!)}{\sum_{i=1}^{N} a_i} \tag{4.9}$$

Figure 4.5 shows that this analytical model of average-case computational complexity does not differ much from the simulated results for N in the range $[4,7]$ and $k = N-1$, i.e., the maximum possible anonymity. It is found that the complexity increases exponentially with N.

4.4 Implementation Issues

In this section, we discuss two practical aspects of the proposed subset coding schemes that need to be resolved in order to implement the scheme in a practical scenario. At first in Section 4.4.1 we consider the temporal fluctuation of attributes and how they may impact the proposed scheme. Then, in Section 4.4.2 another real-life issue of handling non-unique attributes, is explored.

4.4.1 Temporal Fluctuation of Attributes

Attributes of the observed POIs vary over time. In fact, it is this fluctuation that leads to the need for participatory sensing in the first place. The frequency of attribute changes, however, is more or less fixed depending on the nature of the attribute as well as POIs. If sufficient information can be collected, an efficient statistical model can be developed to find the expected period Γ_C between successive changes in attributes among N POIs. As N is considered small in the range [4,6], Γ_C can be estimated to be fairly large. For example, in the *PetrolWatch* application, the price of fuel normally changes once a day and with $N = 6$, it can be easily shown that $\Gamma_C = 24/6 = 4$ hours.

The decoding approach considers all the observations from a reference point beyond which the system has no interest. This approach can be easily modified to incorporate a user-defined temporal window Γ_T, which is preferably set at $\Gamma_\tau \leq \Gamma_C$. The decoder can use the time_of_observation field of each RApS to decide whether it should be included within Γ_τ.

As we aim for high data integrity, Γ_τ also has a lower bound to achieve a desired level of full decodability. For example, with $N = 6$ and $k = 5$, 90% or higher full decodability is achieved with 32 observations, as observed in the simulation result presented in Section 4.5. In this case, Γ_τ must be large enough to accumulate at least 32 RApS. With $n = 100$ registered users, on average, $\omega = 32/100 \approx 0.3$ fraction of users needs to observe during the temporal window. In the previous chapter, we have suggested that ω is fixed, which is governed by the socio-economic and cultural behaviours of the users. In case, ω is smaller than the required value, we can either reduce k, or more promisingly, we can expect that more users will participate as the location privacy improves with the application of the proposed scheme. This participation will effectively keep the required value of ω in check.

We now need to analyse the transient effect of attribute change on decodability, i.e., data integrity. As soon as the changed attribute is reported, it is likely to cause contradictions

with previous reports from the corresponding POI. It is, however, impossible for the ApS to pinpoint the POI due to subset coding with k-anonymity. Consequently, it is likely that the decodability performance of the decoder is degraded due to the contradictions and returns to normalcy after the recovery period Γ_R. We are interested in empirically ascertaining the expected length of Γ_R. In Section 4.5.54.5.6, we have presented the simulation results to conclude that $\Gamma_R \leq \Gamma_\tau$ in all cases of N and k, while aiming to achieve full decodability in 90% of cases or more.

4.4.2 Non-unique Attributes

So far in this chapter we have assumed that the attributes of N POIs are unique, which is unrealistic as POIs are non-communicating. Now we demonstrate that such an assumption keeps the decoding algorithm's complexity in check while the non-unique scenario can be easily transformed to the unique scenario.

When the attributes are assumed to be unique, the decoding algorithm (Algorithm 4.1) needs to consider $N!$ permutations of N-tuples to find the set of tuples conforming all the ARs, resulting in computational complexity of $O(N!)$. If the attributes are non-unique, this algorithm needs to be modified to consider N^N tuples from the N-ary Cartesian product of a set of distinct attributes, increasing computational complexity to $O(N\hat{}N)$. For N, as little as 6, computational complexity of the decoder is increased by $6\hat{}6/6! \approx 65$ times.

The transformation of the non-unique scenario to the unique scenario is quite straightforward. When AS receives an observation report with the attribute that is the same as an already observed attribute of another POI, the AS can make the attribute unique by adding a small value below the level of significance. For example, the price of fuel is normally mentioned in 2 decimal point precision and hence, any value smaller than the unit precision may be used to keep the prices distinct when reported to the ApS.

4.5 Performance Evaluation

In this section, we present simulation results to validate a number of key comparative analyses. All the results presented here were obtained by averaging 1000 simulation runs. Section 4.5.1 presents the simulation setup followed by data integrity performance comparison in Section 4.5.2. Then, in Section 4.5.3 some experimental results are presented to be considered as a reason behind choosing local search direction followed by exploring

Table 4.4: Price Retrieval Rate

k	Approaches	Mean Price Retrieval Rate		
		$N = 3$	$N = 4$	$N = 5$
2	Randomized	83.22%	90.57%	93.40%
	BGAS	89.40%	92.02%	93.51%
	H&K	50.00%	50.00%	50.00%
3	Randomized		65.85%	86.95%
	BGAS		75.34%	93.57%
	H&K		33.33%	33.33%
4	Randomized			40.00%
	BGAS			48.40%
	H&K			25.00%

joint anonymization in Section 4.5.4. Section 4.5.5 presents comparative performance analysis of different techniques discussed so far. Finally, in Section 4.5.6 the impact of the actual fluctuation of attributes is explored in our simulation results.

4.5.1 Simulation Setup

In our simulated participatory sensing system, MNs report observations to AS in a random fashion. The number of POIs (N) were varied from 4 to 6. Since the system is designed in such a way that every locality will have multiple ASs dealing with exclusive POIs, a large number of the same types of POIs are not practical and 4 to 6 was quite realistic. The proposed schemes flexibly handle degree of anonymity requirement (k) and thus results are produced for $k = N - 2$ to $N - 1$. Lower k values imply a lower anonymity which is weak in consideration of privacy. Similarly, we are mostly interested to evaluate decodability performance with a very high degree of data integrity. Therefore, we considered only N-decodability (full decodability) and $(N - 1)$-decodability, termed as partial decodability.

4.5.2 Data Integrity Performance Comparison

Table 4.4 presents the performance of our randomized approach, and BGAS to analyse their performance and compare them with another state-of-the-art, regarding the successful association of the POI-attribute. For a convenient representation of the mean values the unit of measurement was in percentage. The results presented here were for $M = Nk$ length of set

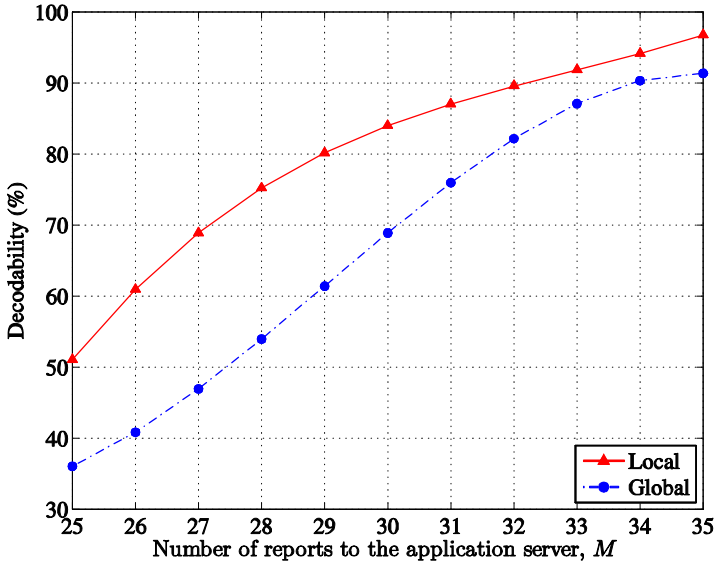


Figure 4.6: Local vs global optima objective in local search: data integrity of ARs in terms of full decodability generated from M number observations when $N = 6$ and $k = 5$.

of anonymized rules. For $N = 5, k = 2$, the achieved integrity 93.51% implies that among the 1000 simulation runs, most of the times correct POI- attribute association was possible for more than 4 out of 5 POIs. Data integrity of Huang *et. al.* [36] was theoretically derived and shown here as H&K approach. As k POIs are reported by a single point (the centre of their tile or alternatively the mean of the equivalence classes), considering reports containing each POI, are same in number, only $1/k$ POI was correctly reported. Table 4.4 clearly presents that in all cases our proposed BGAS approach outperformed both the H&K and randomized one. Hence, from now on, we concentrate on improving our proposed approach and compare with each other.

4.5.3 Local Search Direction

Figure 4.6 presents the data integrity performance of the proposed anonymization algorithm (Algorithm 4.2 vs. Algorithm 4.3) when the local search direction was guided by local (used in EGAS) and global (used in BGAS) optimal objectives for $N = 6, k = 5$. The former achieved full decodability in 90% cases with 2 observations less than needed by the later. This may not seem significant, but considering the level of voluntary participation, any improvement is worth pursuing.

Table 4.5: Full Decodability (%) for Various Degree of Joint Anonymization

M	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
25	52	50	48
26	61	59	57
27	69	67	64
28	76	74	72
29	81	79	78
30	85	83	83
31	88	87	87
32	91	90	91
33	92	93	93
34	94	96	95
35	95	97	96

4.5.4 Joint Anonymization

Table 4.5 presents data integrity performance of the proposed anonymization algorithm when observations are anonymized jointly in groups of $\lambda \in \{1,2,3\}$. It can be observed as predicted in Section 4.3.2.3 that decodability performances for different values of λ are almost indistinguishable.

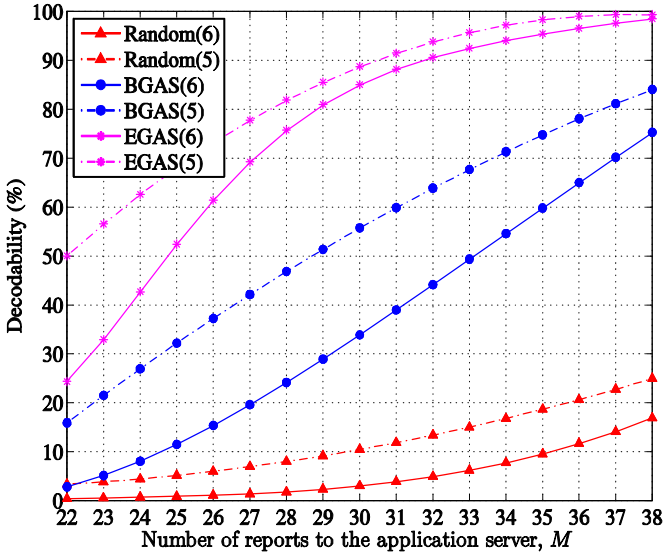
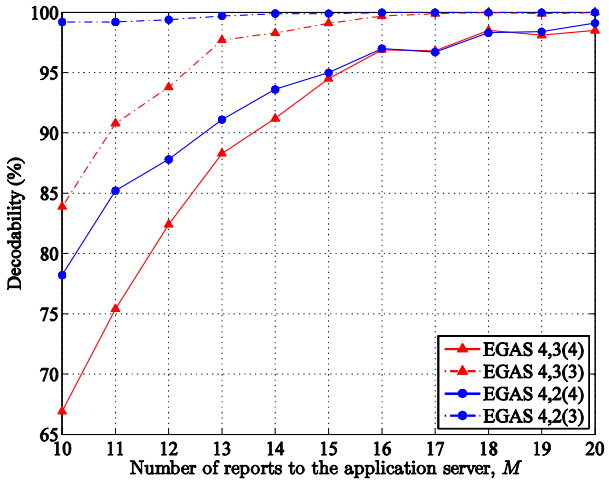
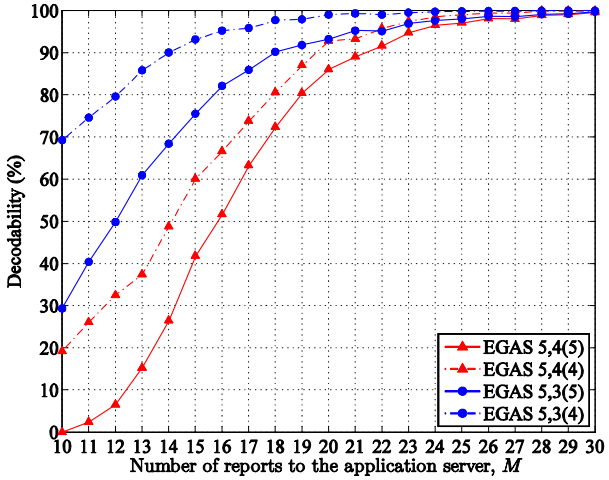


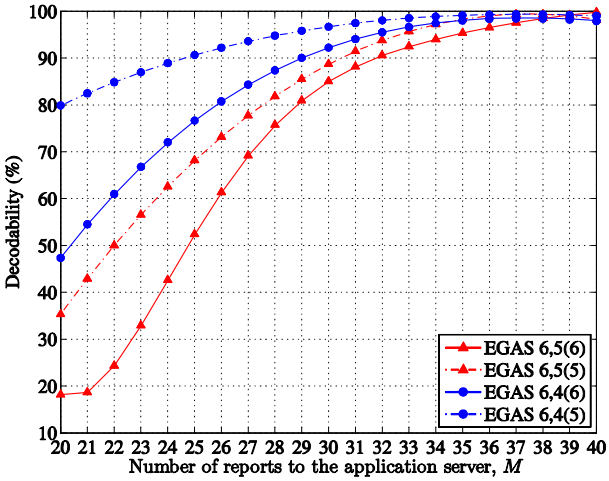
Figure 4.7: Data integrity trend of ARs for different techniques in terms of full (6) and partial (5) decodability generated from M number of observation reports when $N = 6$ and $k = 5$.



(a)



(b)



(c)

Figure 4.8: Data integrity of ARs in terms of full (N) and partial ($N-1$) decodability generated from M number of observation reports when (a) $N = 4$, (b) $N = 5$, and (c) $N = 6$ and in all cases $k = N-1$ and $N-2$.

4.5.5 Comparison among Different Techniques

We have implemented BGAS, EGAS, and a random anonymization scheme in a custom simulator to compare decodability performance for the different number of messages reported by the MNs. In the random anonymization scheme, AS avoids running the decoder by selecting one of the possible ARs at random. In fact, this random scheme does not require the AS as the observing MN can use this algorithm to generate AR from the observation.

Figure 4.7 presents the comparative performance of our proposed EGAS using trend-lines with other two techniques when $N = 6, k = 5$. Trend-lines were produced from regression analysis with a polynomial curve fitting. The degree of polynomial coefficients used for curve fitting were 3 (2), 4 (2), and 6 (3) for full (partial) decodability in case of random, BGAS, and EGAS, respectively. For each scheme, the performance curves for full and partial decodability were plotted with a solid and dashed line, respectively. In some application scenarios, it may be sufficient to have less than full decodability with a requirement of fewer M.

The figure indicates that for a low decodability requirement, $(N - 1)$-decodability may be achieved with significantly fewer observations. In the case of both full and partial decodability, EGAS is superior to the other two techniques. Full decodability was achieved in 90% of simulation runs by EGAS with only 31 observations. For the same number of observations, BGAS attained full decidability in only 42% cases and the random scheme barely (3%) achieved full decodability.

Figure 4.8 shows how D-decodability performance varies for (a) $N = 4$, (b) $N = 5$, and (c) $N = 6$ with $k = N - 1$ and $N - 2$. The performance curve for each set up is identified as $N, k, (D)$ where $D \in \{N, N - 1\}$. In all cases, a certain proportion of the simulation runs

Table 4.6: Number of Observations Needed to Regain 90% Full Decodability after the Attribute of a Randomly Selected POI is Changed

N	k	Γ_T	Γ_R
4	2	13	11
4	3	14	14
5	3	19	17
5	4	22	21
6	4	30	28
6	5	33	33

achieved D-decodability with fewer observation reports when the degree of anonymity (k) or D is lowered. However, to guarantee that in almost all cases the desired decodability is achieved, the minimum number of necessary reports do not vary much with k. This indicates that when a very high degree of decodability is desired, the system can also afford the highest degree of anonymity, i.e., $k = N - 1$, without demanding any significant increase in the number of observations.

4.5.6 Fluctuation of Attributes

In this simulation, we considered the temporal window size Γ_T starting from the number of observations that can achieve 90% full decodability for various N and k, as reported in Figure 4.8. For each window size, the attribute of one of the randomly selected POI was changed when the data integrity was at 90% full decodability or higher. Since that point, we traced the number of new observations needed to recover the data integrity to its original state. In all cases, we observed that the recovery period $\Gamma_R \leq \Gamma_\tau$. Table 4.6 presents the recovery period observed when Γ_τ was minimally set to achieve 90% full decodability.

4.6 Conclusion

In this chapter we presented two *k*-anonymization schemes using the novel subset-coding based anonymization approach introduced in Chapter 3 to protect the privacy of participants in a PSS application. The primary goal in using these techniques was to achieve high quality data after the de-anonymization at the target end. First, we presented our initial approach, BGAS. Next, we discussed the optimization issues needed to improve the quality of BGAS which were successfully addressed and a better scheme EGAS was designed. Performance of both these approaches was compared and also their significant superiority against random anonymization was also established through comprehensive simulation. Thus, the challenging problem of achieving the location privacy of the participating users and, at the same time, the desired data quality at the target end was successfully solved.

However, we feel the need for even better anonymization approaches that can work with fewer observations in such application scenarios. For example, around 40 observations are needed for high data integrity with $N = 6$ indicating that there is ample need for improvement in this regard. Moreover, reducing the computational complexity of the technique would also be a significant improvement. We aim to address these issues in the next chapter.

Chapter 5

5 Efficient Anonymization with Deterministic Techniques

In the previous chapter, subset-coding based anonymization techniques were proposed that attempted to achieve high data integrity at the desired end. We also observed that the necessity of the number of observations from various POIs was quite high to achieve comparable data quality. From that point of view, we aim to anonymize in a way to achieve full decodability with significantly fewer observations. With the existing reward facilitating schemes introduced in PSS, it is quite likely that the required number of observations to achieve full decodability will be collected. From that point on our ApS can provide service deterministically as the system is expected to operate in the Singular Decoding Decision Region shown in Figure 4.1. Reducing computational complexity is another target for these techniques since it would be beneficial in scenarios with a large number of POIs.

The rest of this chapter is organised as follows. In Section 5.1 we introduce the new anonymization technique that follows a deterministic approach. The detail of this basic deterministic approach is presented in Section 5.2 along with formal algorithms. Section 5.3 discusses its fast variation with algorithms related to the proposed approach and an important implementation issue is addressed in Section 5.4. Section 5.5 presents the simulated performance of the proposed schemes and finally Section 5.6 concludes the chapter.

5.1 Introduction

Probabilistic Greedy Anonymization Schemes (PGAS) presented in the previous chapter were simulated in the context of PSS. A clear observation was made that those techniques

would require a good number of observations to provide quality data. This prohibits the applicability of those techniques in applications where some of the POIs are not likely to be observed frequently by the community people due to their remote location.

In this chapter we present new anonymization and decoding schemes that overcome this limitation. The basis of the new schemes is the same subset-coding technique and also the system entities of PSS would remain unchanged. However, now the new approaches philosophically aim to achieve deterministic decoding. Majority decoding was the determinant of the choice of a k-subset at every step in PGAS. Now, the subset that can rule out the maximum number of permutations would be preferred in each step of anonymization. Considering the differences with PGAS, we have named the new techniques as *Deterministic Greedy Anonymization Schemes* (DGAS).

Even DGAS has a limitation of high computational complexity. Although, the PSS can be designed with a number of subsystems each containing a feasible number of POIs, it is still desired to reduce the complexity. Achieving this would offer more flexibility in design, especially in densely populated areas. From this motivation, we have designed an efficient variation of DGAS where choice of k has been restricted to the maximum possible anonymity, $N-1$ only. However, this is satisfactory since this implies the strongest privacy to the participants. To signify the efficient computation of this variation, we have named it

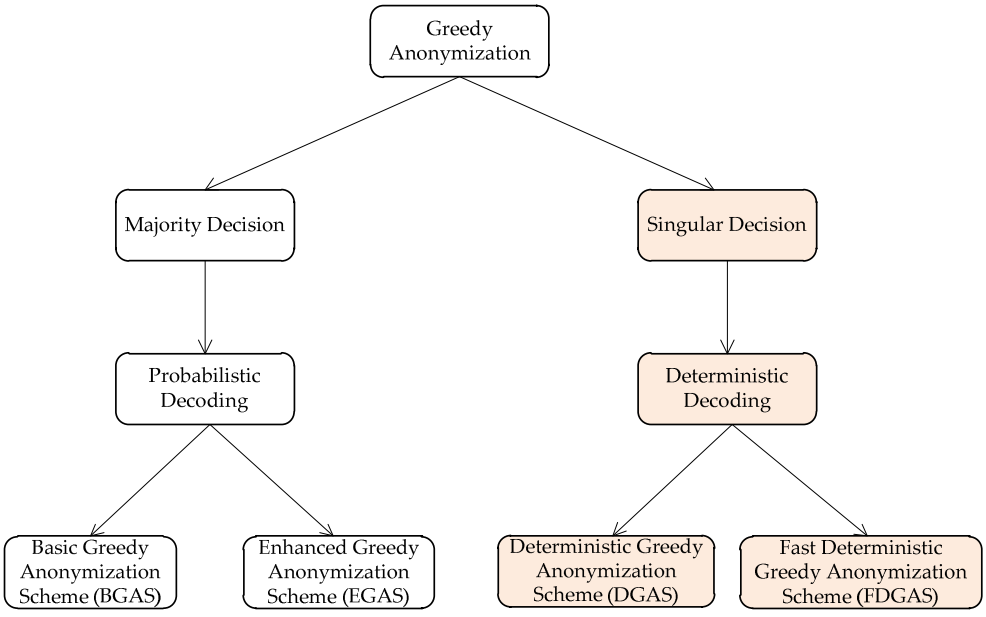


Figure 5.1: Two different approaches of proposed greedy techniques.

Fast Deterministic Greedy Anonymization Schemes (FDGAS). Figure 5.1 highlights the specific branch of anonymization techniques covered in this chapter.

5.2 DGAS

In this section, we first discuss the concept of DGAS in Section 5.2.1. Note that the terms and definitions presented in the previous chapters are used here without any change in meaning. Next, we formally present the anonymization and de-anonymization algorithms in Section 5.2.2.

5.2.1 Concept of DGAS

First of all, we are going to discuss the process by which AS generates ARs from received observations. As we have already mentioned, the approach is a greedy one that tries to augment POI-attribute association with generation of new AR considering the already generated ARs.

As the ORs arrive at AS, our greedy anonymization scheme aims to choose an AR that will maximally reduce the size of $CAAS_{ARS}$. The ultimate target of our proposed scheme is to attain a single cardinality $CAAS_{ARS}$. So, the main objective can be stated as follows.

Objective 5.1 *Achieving singularity, i.e.,* $|CAAS_{ARS}| = 1$ *with minimal observation.*

The AS keeps the POI-Attribute assignment adding some anonymity on it upon getting various ORs from MNs. At the fresh start, when the AS has no information, i.e., no OR has yet been reported to AS, the cardinality of $PAAS$ equals to $N!$. As the ORs start arriving, the AS starts giving suggestion in a way to achieve the objective of reaching singular cardinality $CAAS_{ARS}$. Keeping that objective in mind we develop our greedy algorithm to find corresponding AR that reduces $CAAS_{ARS}$ cardinality maximally. Whenever AS gets an OR, it checks the all possible ARs for it and selects the one which will reduce $CAAS_{ARS}$ cardinality maximally.

Now we will observe a simple example how ARs are selected corresponding to received observations as performed by the AS. Consider an example of *PetrolWatch* with four POIs of ids 1,2,3, and 4, having prices \$10, \$20, \$30, and \$40, respectively. In this $N = 4$ system, the initial size of $PAAS$ is $N! = 24$.

Table 5.1: Conforming Tuples of Gradually Generated AR Set where Dummy Attributes are Identified with Leading d

	1st observation [1, \$10]			ARS
	{1,2,3}: \$10	{1,2,4}: \$10	{1,3,4}: \$10	
$PAAS$:	$(\$10, d1, d2, d3)$	$(\$10, d1, d2, d3)$	$(\$10, d1, d2, d3)$	
$(d1, d2, d3, d4)$	$(\$10, d1, d3, d2)$	$(\$10, d1, d3, d2)$	$(\$10, d1, d3, d2)$	
$(d1, d2, d4, d3)$	$(\$10, d2, d1, d3)$	$(\$10, d2, d1, d3)$	$(\$10, d2, d1, d3)$	
$(d1, d3, d2, d4)$	$(\$10, d2, d3, d1)$	$(\$10, d2, d3, d1)$	$(\$10, d2, d3, d1)$	
$(d1, d3, d4, d2)$	$(\$10, d3, d1, d2)$	$(\$10, d3, d1, d2)$	$(\$10, d3, d1, d2)$	
$(d1, d4, d2, d3)$	$(\$10, d3, d2, d1)$	$(\$10, d3, d2, d1)$	$(\$10, d3, d2, d1)$	
$(d1, d4, d3, d2)$	$(d1, \$10, d2, d3)$	$(d1, \$10, d2, d3)$	~~$(d1, \$10, d2, d3)$~~	
$(d2, d1, d3, d4)$	$(d1, \$10, d3, d2)$	$(d1, \$10, d3, d2)$	~~$(d1, \$10, d3, d2)$~~	
$(d2, d1, d4, d3)$	$(d1, d2, \$10, d3)$	~~$(d1, d2, \$10, d3)$~~	$(d1, d2, \$10, d3)$	
$(d2, d3, d1, d4)$	~~$(d1, d2, d3, \$10)$~~	$(d1, d2, d3, \$10)$	$(d1, d2, d3, \$10)$	
$(d2, d3, d4, d1)$	$(d1, d3, \$10, d2)$	~~$(d1, d3, \$10, d2)$~~	$(d1, d3, \$10, d2)$	
$(d2, d4, d1, d3)$	~~$(d1, d3, d2, \$10)$~~	$(d1, d3, d2, \$10)$	$(d1, d3, d2, \$10)$	
$(d2, d4, d3, d1)$	$(d2, \$10, d1, d3)$	$(d2, \$10, d1, d3)$	~~$(d2, \$10, d1, d3)$~~	
$(d3, d1, d2, d4)$	$(d2, \$10, d3, d1)$	$(d2, \$10, d3, d1)$	~~$(d2, \$10, d3, d1)$~~	
$(d3, d1, d4, d2)$	$(d2, d1, \$10, d3)$	~~$(d2, d1, \$10, d3)$~~	$(d2, d1, \$10, d3)$	
$(d3, d2, d1, d4)$	~~$(d2, d1, d3, \$10)$~~	$(d2, d1, d3, \$10)$	$(d2, d1, d3, \$10)$	
$(d3, d2, d4, d1)$	$(d2, d3, \$10, d1)$	~~$(d2, d3, \$10, d1)$~~	$(d2, d3, \$10, d1)$	
$(d3, d4, d1, d2)$	~~$(d2, d3, d1, \$10)$~~	$(d2, d3, d1, \$10)$	$(d2, d3, d1, \$10)$	
$(d3, d4, d2, d1)$	$(d3, \$10, d1, d2)$	$(d3, \$10, d1, d2)$	~~$(d3, \$10, d1, d2)$~~	
$(d4, d1, d2, d3)$	$(d3, \$10, d2, d1)$	$(d3, \$10, d2, d1)$	~~$(d3, \$10, d2, d1)$~~	
$(d4, d1, d3, d2)$	$(d3, d1, \$10, d2)$	~~$(d3, d1, \$10, d2)$~~	$(d3, d1, \$10, d2)$	
$(d4, d2, d1, d3)$	~~$(d3, d1, d2, \$10)$~~	$(d3, d1, d2, \$10)$	$(d3, d1, d2, \$10)$	
$(d4, d2, d3, d1)$	$(d3, d2, \$10, d1)$	~~$(d3, d2, \$10, d1)$~~	$(d3, d2, \$10, d1)$	
$(d4, d3, d1, d2)$	~~$(d3, d2, d1, \$10)$~~	$(d3, d2, d1, \$10)$	$(d3, d2, d1, \$10)$	
$(d4, d3, d2, d1)$				
CR	6	6	6	
$CAAS_{ARS}$:	2nd observation [3, \$30]			{1,2,3}: \$10
$(\$10, d1, d2, d3)$	{1,2,3}: \$30	{1,3,4}: \$30	{2,3,4}: \$30	
$(\$10, d1, d3, d2)$	$(\$10, \$30, d1, d2)$	~~$(\$10, \$30, d1, d2)$~~	$(\$10, \$30, d1, d2)$	
$(\$10, d2, d1, d3)$	$(\$10, \$30, d2, d1)$	~~$(\$10, \$30, d2, d1)$~~	$(\$10, \$30, d2, d1)$	
$(\$10, d2, d3, d1)$	$(\$10, d1, \$30, d2)$	$(\$10, d1, \$30, d2)$	$(\$10, d1, \$30, d2)$	
$(\$10, d3, d1, d2)$	~~$(\$10, d1, d2, \$30)$~~	$(\$10, d1, d2, \$30)$	$(\$10, d1, d2, \$30)$	
$(\$10, d3, d2, d1)$	$(\$10, d2, \$30, d1)$	$(\$10, d2, \$30, d1)$	$(\$10, d2, \$30, d1)$	
$(d1, \$10, d2, d3)$	~~$(\$10, d2, d1, \$30)$~~	$(\$10, d2, d1, \$30)$	$(\$10, d2, d1, \$30)$	
$(d1, \$10, d3, d2)$	$(\$30, \$10, d1, d2)$	$(\$30, \$10, d1, d2)$	~~$(\$30, \$10, d1, d2)$~~	
$(d1, d2, \$10, d3)$	$(\$30, \$10, d2, d1)$	$(\$30, \$10, d2, d1)$	~~$(\$30, \$10, d2, d1)$~~	
$(d1, d3, \$10, d2)$	$(\$30, d1, \$10, d2)$	$(\$30, d1, \$10, d2)$	~~$(\$30, d1, \$10, d2)$~~	
$(d2, \$10, d1, d3)$	$(\$30, d2, \$10, d1)$	$(\$30, d2, \$10, d1)$	~~$(\$30, d2, \$10, d1)$~~	
$(d2, \$10, d3, d1)$	$(d1, \$10, \$30, d2)$	$(d1, \$10, \$30, d2)$	$(d1, \$10, \$30, d2)$	
$(d2, d1, \$10, d3)$	~~$(d1, \$10, d2, \$30)$~~	$(d1, \$10, d2, \$30)$	$(d1, \$10, d2, \$30)$	
$(d2, d3, \$10, d1)$	$(d1, \$30, \$10, d2)$	~~$(d1, \$30, \$10, d2)$~~	$(d1, \$30, \$10, d2)$	
$(d3, \$10, d1, d2)$	~~$(d1, d2, \$10, \$30)$~~	$(d1, d2, \$10, \$30)$	$(d1, d2, \$10, \$30)$	
$(d3, \$10, d2, d1)$	$(d2, \$10, \$30, d1)$	$(d2, \$10, \$30, d1)$	$(d2, \$10, \$30, d1)$	
$(d3, d1, \$10, d2)$	~~$(d2, \$10, d1, \$30)$~~	$(d2, \$10, d1, \$30)$	$(d2, \$10, d1, \$30)$	
$(d3, d2, \$10, d1)$	$(d2, \$30, \$10, d1)$	~~$(d2, \$30, \$10, d1)$~~	$(d2, \$30, \$10, d1)$	
	~~$(d2, d1, \$10, \$30)$~~	$(d2, d1, \$10, \$30)$	$(d2, d1, \$10, \$30)$	
CR	6	4	4	

$CAAS_{ARS}$:	3rd observation [2, $20]			{1,2,3}: $10
($10, $30, $d1, $d2)	{1,2,3}: $20	{1,2,4}: $20	{2,3,4}: $20	{1,2,3}: $30
($10, $30, $d2, $d1)	($10, $30, $20, $d1)	~~($10, $30, $20, $d1)~~	($10, $30, $20, $d1)	
($10, $d1, $30, $d2)	~~($10, $30, $d1, $20)~~	($10, $30, $d1, $20)	($10, $30, $d1, $20)	
($10, $d2, $30, $d1)	($10, $20, $30, $d1)	($10, $20, $30, $d1)	($10, $20, $30, $d1)	
($30, $10, $d1, $d2)	~~($10, $d1, $30, $20)~~	($10, $d1, $30, $20)	($10, $d1, $30, $20)	
($30, $10, $d2, $d1)	($30, $10, $20, $d1)	~~($30, $10, $20, $d1)~~	($30, $10, $20, $d1)	
($30, $d1, $10, $d2)	~~($30, $10, $d1, $20)~~	($30, $10, $d1, $20)	($30, $10, $d1, $20)	
($30, $d2, $10, $d1)	($30, $20, $10, $d1)	($30, $20, $10, $d1)	($30, $20, $10, $d1)	
($d1, $10, $30, $d2)	~~($30, $d1, $10, $20)~~	($30, $d1, $10, $20)	($30, $d1, $10, $20)	
($d1, $30, $10, $d2)	($20, $10, $30, $d1)	($20, $10, $30, $d1)	~~($20, $10, $30, $d1)~~	
($d2, $10, $30, $d1)	($20, $30, $10, $d1)	($20, $30, $10, $d1)	~~($20, $30, $10, $d1)~~	
($d2, $30, $10, $d1)	~~($d1, $10, $30, $20)~~	($d1, $10, $30, $20)	($d1, $10, $30, $20)	
	~~($d1, $30, $10, $20)~~	($d1, $30, $10, $20)	($d1, $30, $10, $20)	
CR	6	2	2	

$CAAS_{ARS}$:	4th observation [4, $40]			{1,2,3}: $10
($10, $30, $20, $d1)	{1,2,4}: $40	{1,3,4}: $40	{2,3,4}: $40	{1,2,3}: $30
($10, $20, $30, $d1)	($10, $30, $20, $40)	($10, $30, $20, $40)	($10, $30, $20, $40)	{1,2,3}: $20
($30, $10, $20, $d1)	($10, $20, $30, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	
($30, $20, $10, $d1)	($30, $10, $20, $40)	($30, $10, $20, $40)	($30, $10, $20, $40)	
($20, $10, $30, $d1)	($30, $20, $10, $40)	($30, $20, $10, $40)	($30, $20, $10, $40)	
($20, $30, $10, $d1)	($20, $10, $30, $40)	($20, $10, $30, $40)	($20, $10, $30, $40)	
	($20, $30, $10, $40)	($20, $30, $10, $40)	($20, $30, $10, $40)	
CR	0	0	0	

$CAAS_{ARS}$:	5th observation [2, $20]			{1,2,3}: $10
($10, $30, $20, $40)	{1,2,3}: $20	{1,2,4}: $20	{2,3,4}: $20	{1,2,3}: $30
($10, $20, $30, $40)	($10, $30, $20, $40)	~~($10, $30, $20, $40)~~	($10, $30, $20, $40)	{1,2,3}: $20
($30, $10, $20, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	{1,2,4}: $40
($30, $20, $10, $40)	($30, $10, $20, $40)	~~($30, $10, $20, $40)~~	($30, $10, $20, $40)	
($20, $10, $30, $40)	($30, $20, $10, $40)	($30, $20, $10, $40)	($30, $20, $10, $40)	
($20, $30, $10, $40)	($20, $10, $30, $40)	($20, $10, $30, $40)	~~($20, $10, $30, $40)~~	
	($20, $30, $10, $40)	($20, $30, $10, $40)	~~($20, $30, $10, $40)~~	
CR	0	2	2	

$CAAS_{ARS}$:	6th observation [2, $20]			{1,2,3}: $10
($10, $20, $30, $40)	{1,2,3}: 20	{1,2,4}: $20	{2,3,4}: $20	{1,2,3}: $30
($30, $20, $10, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	{1,2,3}: $20
($20, $10, $30, $40)	($30, $20, $10, $40)	($30, $20, $10, $40)	($30, $20, $10, $40)	{1,2,4}: $40
($20, $30, $10, $40)	($20, $10, $30, $40)	($20, $10, $30, $40)	~~($20, $10, $30, $40)~~	{1,2,4}: $20
	($20, $30, $10, $40)	($20, $30, $10, $40)	~~($20, $30, $10, $40)~~	
CR	0	0	2	

$CAAS_{ARS}$:	7th observation [1, $10]			{1,2,3}: $10
($10, $20, $30, $40)	{1,2,3}: $10	{1,2,4}: $10	{1,3,4}: $10	{1,2,3}: $30
($20, $10, $30, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	($10, $20, $30, $40)	{1,2,3}: $20
	($30, $20, $10, $40)	~~($30, $20, $10, $40)~~	($30, $20, $10, $40)	{1,2,4}: $40
CR	0	1	0	{1,2,4}: $20
				{2,3,4}: $20

$CAAS_{ARS}$:	{1,2,3}: $10
($10, $20, $30, $40)	{1,2,3}: $30
	{1,2,3}: $20
	{1,2,4}: $40
	{1,2,4}: $20
	{2,3,4}: $20
	{1,2,4}: $10

Table 5.1 demonstrates how the ARS is generated and populated by our greedy approach so that single cardinality $CAAS_{ARS}$ is achieved, i.e., all four POIs can be associated with the correct price for a minimal set of observation. Here, the updated $CAAS_{ARS}$ is shown each time an AR is selected for each observation.

When the first OR $[1, \$10]$ is received by the AS, it uses three dummy prices. For desired anonymity $k = 3$, there are three possible ARs that can be ranked based on their ability of $CAAS_{ARS}$ cardinality reduction. Cardinality Reduction (CR) means count of eliminating non-conforming permutations after applying an AR. As any of the three possible ARs, $\{1,2,3\}: \$10$, $\{1,2,4\}: \$10$, and $\{1,3,4\}: \$10$, can reduce $CAAS_{ARS}$ cardinality by 6, we can pick any. Let us assume that we pick $\{1,2,3\}: \$10$ for the first observation, after which cardinality of $CAAS_{ARS}$ reduces to $N! - 6 = 18$.

When the second OR $[3, \$30]$ arrives, the AS needs to use two dummy prices. Among the three possible ARs, $\{1,2,3\}: \$30$ can reduce cardinality by 6, whereas both the other two $\{1,3,4\}: \$30$ and $\{2,3,4\}: \$30$ can reduce that by 2. Hence, the AS selects the AR $\{1,2,3\}: \$30$ in response to this observed report and after applying that the cardinality of $CAAS_{ARS}$ reduces

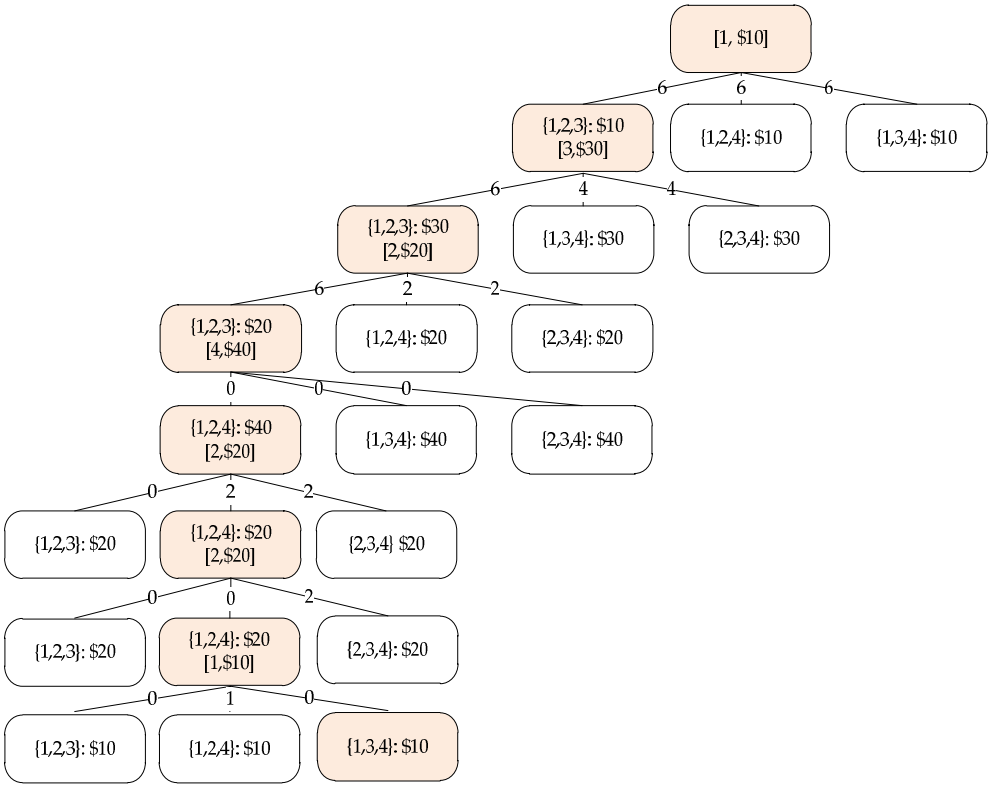


Figure 5.2: Subset generation procedure in DGAS.

to $18 - 6 = 12$.

When the third OR $[2, \$20]$ is received, AS now needs to use only one dummy price. Of the three possible ARs, $\{1,2,3\}: \$20$ can reduce cardinality by 6, whereas the other two $\{1,2,4\}: \$20$ and $\{2,3,4\}: \$20$ score 2 in CR. Therefore, the AR $\{1,2,3\}: \$20$ with maximum CR score is picked in response to this observation. Then, the cardinality of the updated $CAAS_{ARS}$ becomes $12 - 6 = 6$.

When the fourth observation $[4, \$40]$ arrives, none of the possible ARs can reduce cardinality any more. Hence, AS randomly picks an AR, $\{1,2,4\}: \$40$, after applying which $CAAS_{ARS}$ cardinality remains the same as before, i.e., 6. Finally, after seven observations, it reaches the objective of achieving $CAAS_{ARS}$ cardinality 1.

Note from the table that $CAAS_{ARS}$ cardinality gradually decreases with the number of ARs. By adding a new AR to the set ARS, some of the existing conforming tuples no longer remain conforming to the new set of ARs. Effectively, only the tuple representing correct POI-attribute association for all POIs survives as the single cardinality $CAAS_{ARS}$.

For the sake of completeness, the same example is shown in a tree-format in Figure 5.2 where one-by-one seven ARs are generated in the AS with seven incoming ORs. The solid edges of the tree refer to the number of non-conforming permutations elimination, i.e., CR score after applying an AR for the given OR. It is clear that at each step the AR with maximum CR score is selected to be appended with the next level subset derived from the next input. In the case of more than one subset exhibiting a maximum CR score, any one is chosen to be the selected subset.

While decoding at the ApS end, it starts with the initial $PAAS$ with cardinality $N!$. Upon receiving each RApS, it cancels out the non-conforming tuples from the $PAAS$. As the same as anonymization at AS, decoding at ApS also aims to achieve cardinality 1. However, in some cases decoder at ApS may arrive at $CAAS_{ARS}$ cardinality equal to zero. It means a contradictory RApS has been received and taken into account by the decoder of ApS that has led to reach $CAAS_{ARS}$ cardinality 0. Here, we can state a corollary as follows.

Corollary 5.2: *Emptiness i.e.,* $|CAAS_{ARS}| = 0$ *implies inconclusive state as a result of either fluctuation or fraudulent.*

This contradiction may arise from genuine fault or from a change in attribute value or from intentional false data feeding to fail the system. Again, when ApS fails to handle data

Algorithm 5.1: AR = **Anonymize_DGAS** $(N, k, \&(\alpha_1, \dots, \alpha_N), \&ARS, \&C, o \equiv [i, a_i])$

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Actual attributes of all POIs from recent observations $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, $\alpha_i$is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values
- Set of already generated ARs, ARS
- Observation report $o \equiv [i, a_i]$, $1 \leq i \leq N$
- Updated $CAAS_{ARS}$, C. Initially, when $ARS = \emptyset$, $|C| = N!$.

Output:

- Anonymized rule $AR \equiv \{i_1, \dots, i_k\}: a_i | i \in \{i_1, \dots, i_k\}$

1. IF $\alpha_i \geq 0$
2. Set $a_i = -a_i$
3. Set $ARS = \emptyset$
4. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
5. Set $C = \{c_1, \dots, c_{N!}\}$ where $c_i = \left(c_{i,1}, \dots, c_{i,N}\right)$ is a unique permutation of $(\alpha_1, \dots, \alpha_N)$ for all i
6. END IF
7. Set $\alpha_i = |a_i|$
8. Set $ss = \underset{\forall ss \in PAS_i^{\{1,\dots,N\},k}}{\operatorname{argmin}} |C \oplus (ss: \alpha_i)|$
9. Set $AR = (ss: a_i)$
10. Set $ARS = ARS \cup \{AR\}$
11. Set $C = C \oplus AR$

from different region differently, this inconclusive state may take place. Now, we formally present the algorithms of DGAS and its efficient variation FDGAS to deterministically k-anonymize the observations by AS and decode those in ApS.

5.2.2 The Anonymization and Decoding Algorithms

As DGAS anonymizes individual observations with the aim of achieving single cardinality $CAAS_{ARS}$, it is essential to measure CR score of the tentative rules. Here, we first present the anonymization scheme (Algorithm 5.1) and then the decoding scheme (Algorithm 5.2).

Algorithm 5.1 removes all past observations when attribute fluctuation is detected, which is then signalled by encoding the attribute with a negative sign (steps 1-6). It records the new attribute in step 7. Finally, it constructs the AR using a subset that removes maximal permutations. Here we have removed the dependency of AS on decoder call to take decision

of selecting a suitable AR, as used in our earlier anonymization algorithms (Algorithms 4.2, 4.3). However, we still need to maintain and update C.

Lemma 5.3. *Computational complexity of Algorithm 5.1 is* $O(N^N)$.

Proof. Let the time complexity of the algorithm is $T(N)$. In line 5, the algorithm takes $O(N!)$ time as $|C| = O(N!)$. Then in line 8, the algorithm checks each subset of PAS and $\left|PAS_i^{\{1,\dots,N\},k}\right| \leq \binom{N-1}{k-1} = O(2^N)$, for $k = \frac{N}{2}$ as it represents the worst-case scenario. Hence the total complexity of the algorithm, $T(N) = O(2^N \times N!) = O(N^N)$. ■

Algorithm 5.2 receives a set of ARs and finds the set of attributes from these ARs. The decoder is reset at the beginning or when the previous decoding contradicted due to a malicious AR or when the AS signals attribute fluctuation or when the number of attributes

Algorithm 5.2: $(da_1, \dots, da_N) =$ **Decode_DGAS**$\left(N, k, \&(\alpha_1, \dots, \alpha_N), \&C, AR \equiv (ss: a)\right)$

Input:

- Number of POIs, N
- Degree of desired anonymity, k
- Actual attributes of all POIs from recent observations $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, $\alpha_i$is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values
- Set of permutations $C = \{c_i\}$ where $c_i = (p_1, \dots, p_N)$ derived from the set $\{1, \dots, N\}$ without contradicting with ARS. Initially, when $ARS = \emptyset$, $|C| = N!$.
- Anonymized Rule, $AR \equiv (ss: a)$

Output:

- Decoded attributes, $(da_1, \dots, da_N)$

1. IF $C = \emptyset \vee a < 0 \vee \left(\forall i: \alpha_i \geq 0 \wedge \nexists j: \alpha_j = a\right)$ THEN
2. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
3. Set $C = \{c_1, \dots, c_{N!}\}$ where $c_i = \left(c_{i,1}, \dots, c_{i,N}\right)$ is a unique permutation of $(\alpha_1, \dots, \alpha_N)$ for all i
4. END IF
5. Set $\alpha_j = |a|$ where $j = \underset{\forall i}{\operatorname{argmin}}\, \alpha_i < 0$
6. Set $C = C \oplus AR$
7. IF $C \neq \emptyset$ THEN
8. FOR $i = 1, \dots, N$ DO
9. Set $da_i = \begin{cases} c_{1,i}, & \text{if } c_{1,i} = \cdots = c_{|C|,i} \geq 0 \\ -\infty, & \text{otherwise} \end{cases}$
10. END FOR
11. ELSE
12. Set $da_i = \infty$, for all $1 \leq i \leq N$
13. END IF

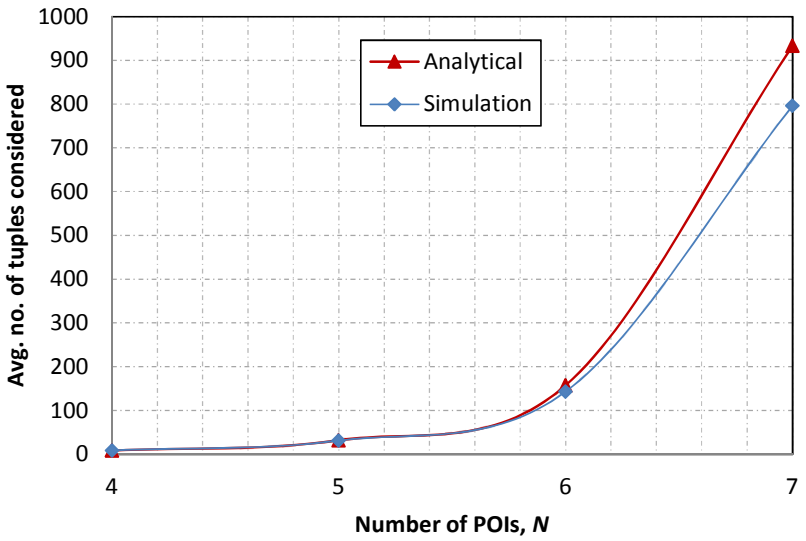


Figure 5.3: Average number of permutations considered each time to check the conformity of an AR varying with N and k.

exceeds N due to malicious AR (steps 1-4). Then, it removes permutations non-conforming to AR after recording the new attribute (steps 5, 6). Finally, it decodes, when deterministically possible or indicates not-yet-decodable (steps 7-10) or declares a contradiction when no permutation survives (steps 11-13).

Lemma 5.4. *Computational complexity of Algorithm 5.2 is* $O(N^N)$.

Proof. Let the time complexity of the algorithm is $T(N)$. In line 3, the algorithm takes $O(N!)$ time as $|C| = O(N!)$. This operation is the dominating one in this algorithm. Hence, the total complexity of the algorithm, $T(N) = O(N!) = O(N^N)$. ■

However, these computational complexities are considered for the worst-case scenario. For the average case scenario we have developed an expression as discussed in the previous chapter in Equation (4.9). Figure 5.3 shows that this theory-based approximation does not differ much from what is acquired using simulation. The graph shows results for N in the range $[4, 7]$ and $k = N - 1$, which is the maximum possible anonymity to compare the average number of tuples considered to check conformity with the AR each time using simulation versus theoretical analysis. It was found that the complexity increases exponentially with N.

5.3 FDGAS

In PGAS, we focused on majority decoding. However, the ultimate target of all time is to provide service with reliability. The sooner we reach single cardinality $CAAS_{ARS}$ the better the purpose is served. From that realization we have developed the philosophy of DGAS and focused on selecting ARs that reduce $CAAS_{ARS}$ cardinality maximally at every stage of anonymization. In doing so, we made an important observation, which in turn helps in developing the concept of a faster approach of this deterministic scheme. In this section, we first discuss the concept of FDGAS in Section 5.3.1. Next, we formally present the anonymization and de-anonymization algorithms in Section 5.3.2.

5.3.1 Concept of FDGAS

In Table 5.1, it is evident that the conforming tuples for any AR $\{i_1, \dots, i_k\}: a_i$ do not have attribute a_i in the POI column, which is not represented in the subset $\{i_1, \dots, i_k\}$. For the first observation $[1, \$10]$, when AR $\{1,2,3\}: \$10$ is considered, none of conforming tuples has \$10 in column 4 and similarly, when AR $\{1,2,4\}: \$10$ is considered, all the tuples having \$10 in column 3 get eliminated. We have now established the following lemma:

Lemma 5.5: *When $k = N - 1$, an AR $\{i_1, \dots, i_{N-1}\}: a_i$ renders all N-tuples having attribute a_i associated with the missing POI in the given subset $\{i_1, \dots, i_{N-1}\}$, i.e., $i_N = \{1, \dots, N\} \setminus \{i_1, \dots, i_k\}$ non-conforming, i.e., these tuples get eliminated.* ■

For an observation of attribute a_i, if one of the maximal AR $\{i_1, \dots, i_{N-1}\}: a_i$ is selected that can eliminate the maximum possible non-conforming tuples, the remaining conforming tuples show an interesting property. For each attribute value in column i_N, the number of its instances in any other column cannot be higher, as those columns may also have instances of attribute a_i, which no longer exists in column i_N. Hence, the following corollary is established:

Corollary 5.6: *When $k = N - 1$, among the conforming N-tuples after selecting a maximal AR $\{i_1, \dots, i_{N-1}\}: a$ that eliminates the maximum possible non-conforming tuples, for each attribute values in column i_N, no other column has more instances of the attribute.* ■

For example, in Table 5.1, the conforming tuples after selecting a maximal AR $\{1,2,3\}: \$10$ for the first observation has the highest instances (6) of [dummy] attribute values d_1, d_2, and d_3 in column $\{1,2,3,4\} \setminus \{1,2,3\} = 4$. Similarly, the conforming tuples after

selecting a maximal AR $\{1,2,4\}$: \$20 for the fifth observation has the highest instances (2) of attribute values \$10 and \$30 in column $\{1,2,3,4\} \setminus \{1,2,4\} = 3$.

When maximal AR is selected for each observation to eliminate maximum possible non-conforming tuples, Corollary 5.6 provides a fast way to select the maximal subset for any future observation, which is expressed in the following theorem:

Theorem 5.7: *When $k = N - 1$, to anonymize an observation $[i, a]$, any subset in $PAS_i^{\{1,\dots,N\},k}$ that has already been used for anonymizing observations from any other POI but i will eliminate maximum possible non-conforming tuples.* ■

A fast, polynomial time bounded, anonymization algorithm springs from this theorem, which is formally defined in the next section. Given a sequence of observations $\langle i_1, i_2, i_3, i_4, i_1 \rangle$ for the corresponding domain of attributes $\langle a_{i_1}, a_{i_2}, a_{i_3}, a_{i_4}, a_{i_1} \rangle$ in a $k = 3$ and $N = 4$ scenario, the initial $PAAS$-cardinality is equal to $N!$. For the very first observation, any possible AR can be selected randomly as all possible ARs result in same cardinality reduction for they are all equally likely at that very first instance. If the AR $\{i_1, i_2, i_3\}: a_{i_1}$ is selected for the first OR, it implies that a_{i_1} cannot be the attribute of the POI i_4. As soon as an attribute is selected to be associated with the subset, the other attributes get equally eliminated with the same statistics for the POIs in the subset but preserved for the non-present one.

After selecting an AR, CAAS-cardinality is updated by eliminating the non-conforming attribute N-tuples and they are fixed for consideration when the next OR arrives. Then after $m = 3$ observations, $CR_m \geq CR'_m$, where CR_m and CR'_m denote counts of cardinality reduction using this theorem and otherwise respectively after m number of observations. That is, for the next 2 observations the same subset $\{i_1, i_2, i_3\}$ can give the best cardinality reduction if selected. For the other possible ARs, some of the non-conforming attribute N-tuples have already got eliminated once the previous AR has been selected. However, upon selecting the previous AR at the first place, the entire non-conforming attribute N-tuples for selecting next AR with the same anonymized subset part representing a different observation are all kept conforming.

Careful scrutiny of Table 5.1 also reveals that column i of the conforming tuples guarantees deterministic decoding of POI i when all the subsets in $PAS_i^{\{1,\dots,N\},k}$ have already been used to anonymize observations from POI i. After anonymizing the sixth observation,

POI 2 has been observed thrice and all the three subsets in $PAS_2^{\{1,\dots,4\},3} = \{\{1,2,3\},\{1,2,4\},\{2,3,4\}\}$ have been used to anonymize the corresponding attribute \$20. Only two 4-tuples have survived and their column 2 has only one value \$20 to guarantee that POI 2 can be decoded correctly.

A recursive fast, polynomial time bounded, decoding algorithm springs from this scrutiny, which is formally defined in the next section. According to this approach, the decoder checks in the ARS whether for any single attribute, a complete PAS for any of the POIs has appeared or not. Once it obtains a complete PAS for a particular POI, then that POI is actually decoded and can now be correctly associated with that single attribute. We then use the same approach recursively to decode another POI in the reduced problem space $(N-1, k-1)$ and so on.

In order to define this recursive approach, the definition of PAS in (4.1) needs to be modified as follows:

$$PAS_i^{I,k} = \begin{cases} \emptyset, & \text{if } k \geq |I| \vee i \notin I \\ \{\{i, i_1, \dots, i_{k-1}\} | \{i_1, \dots, i_{k-1}\} \subset I \setminus \{i\}\}, & \text{otherwise} \end{cases} \quad (5.1)$$

where, I represents the set of yet-to-be-decoded POIs. We have now established the following theorem from the above scrutiny:

Theorem 5.8: *When $k = N-1$, POI i can be decoded deterministically if all subsets in $PAS_i^{I,k-N+|I|}$, augmented with so-far-decoded set of POIs $\{1,\dots,N\} \setminus I$, have already been used to anonymize observations from POI i where I is the set of yet-to-be-decoded POIs.* ■

Note that, this fast approach is valid only for $k = N-1$ which is the best possible privacy preserving case. With extensive experimentation, it has been found that FDGAS gives the same result as DGAS, and only differs in computational complexity level. Hence, it may perform satisfactorily in bigger scenarios with larger values of N.

5.3.2 The Anonymization and Decoding Algorithms

To anonymize with the aim of achieving lower computational complexity, we need to avoid the decoder call from AS as well as avoid maintaining and updating C. On the basis of some observation heuristics, we then design FDGAS for the cases where, $k = N-1$ which now have computational complexity reduced to polynomial order. We first present the anonymization scheme (Algorithm 5.3) and then the decoding scheme (Algorithm 5.4).

Algorithm 5.3 presents the anonymization technique in light of Theorem 5.7 where $k = N - 1$. It removes all past observations when attribute fluctuation is detected, which is then signalled by encoding the attribute with a negative sign (steps 1-5). Then, it records the new attribute in step 6. Finally, it constructs the AR using a subset, preferably non-repeated, that has already been used maximally for anonymizing other attributes (steps 7-14).

Lemma 5.9. *Computational complexity of Algorithm* 5.3 *is* $O(N^3)$.

Proof. Let the time complexity of the algorithm is $T(N)$. Line 8 represents the most dominating computation. In the worst case, the algorithm takes $O(N^2)$ time as $\left|PAS_i^{I,k}\right| \leq \binom{N-1}{N-2} = O(N)$ and $|S| \leq O(N^2)$. Hence the total complexity of the algorithm, $T(N) = O(N \times N^2) = O(N^3)$. ■

Algorithm 5.3: AR = **Anonymize_FDGAS** $(N, k, \&(\alpha_1, \dots, \alpha_N), \&S, o \equiv [i, a_i])$

Input:

- Number of POIs, N
- Degree of desired anonymity, $k = N - 1$
- Actual attributes of all POIs from recent observations $(\alpha_1, \dots, \alpha_N)$; if attribute of any POI i is yet to be observed, α_i is set to its unique dummy value $-i$ assuming that real attributes are always mapped to positive values
- Set of already generated ARs, S
- Observation report $o \equiv [i, a_i]$, $1 \leq i \leq N$

Output:

- Anonymized rule $AR \equiv \{i_1, \dots, i_k\}: a_i | i \in \{i_1, \dots, i_k\}$

1. IF $\alpha_i \geq 0$
2. Set $a_i = -a_i$
3. Set $S = \emptyset$
4. Set $(\alpha_1, \dots, \alpha_N) = (-1, \dots, -N)$
5. END IF
6. Set $\alpha_i = |a_i|$
7. Set $I = \{1, \dots, N\}$
8. Set $SS = \left\{ss \middle| ss \in PAS_i^{I,k} \wedge |\{a|(ss: a) \in S \wedge a \neq \alpha_i\}| = \max_{\forall j \in PAS_i^{I,k}} |\{b|(j: b) \in S \wedge b \neq \alpha_i\}|\right\}$
9. IF $|SS| = 1$ THEN
10. Set $AR = (ss: a_i)$ where $ss \in SS$
11. ELSE
12. Set $AR = (ss: a_i)$ where $ss \in SS \wedge (ss: \alpha_i) \notin S$
13. END IF
14. Set $S = S \cup \{AR\}$

Algorithm 5.4: $(da_1, \dots, da_N) = \textbf{Decode\_FDGAS}\big(N, k \equiv N-1, \&S, AR \equiv (ss{:}\, a)\big)$

Input:

- Number of POIs, N
- Degree of desired anonymity, $k = N - 1$
- Set of already generated ARs, S
- Anonymized Rule, $AR \equiv (ss{:}\, a)$

Output:

- Decoded attributes, $(da_1, \dots, da_N)$

1. IF $a < 0 \vee |\{b | \exists j{:}\,(j{:}\, b) \in S\} \cup \{a\}| > N$ THEN
2. Set $S = \emptyset$
3. END IF
4. Set $S = S \cup \{(ss{:}\, |a|)\}$
5. Set $da_i = -\infty$ for all $1 \leq i \leq N$
6. Set $I = \{1, \dots, N\}$
7. Set $N_D = \textbf{Decode\_FDGAS\_R}(N, k, S, \&(da_1, \dots, da_N), I)$
8. IF $N_D = N - 1 \wedge \exists (j{:}\, b) \in S{:}\, b \notin (da_1, \dots, da_N)$ THEN
9. Set $da_i = b$ for i such that $da_i = -\infty$
10. END IF

Algorithm 5.4 presents the decoding technique in light of Theorem 5.8 where $k = N - 1$. It receives a set of ARs and finds the set of attributes from these ARs. It resets the decoder when the AS signals attribute fluctuation or number of attributes exceeds N due to false data feeding (steps 1-3). Then it appends AR to the collection after changing the attribute sign to positive in step 4. Finally, it decodes up to $N - 1$ attributes using the recursive decoder by assuming not-yet-decodable at the beginning (steps 5-10).

The recursive decoder in Algorithm 5.5 decodes only when the collection is non-empty (step 1). At first, it finds an attribute with maximum number of subsets available in S (steps 2, 3). Then it declares contradiction when number of subsets equals N, which is impossible. Finally, it assigns that attribute to POI i if the subsets represent $PAS_i^{I,k}$ and then recursively decodes the remaining attributes by eliminating the ARs of the selected attribute as well as other ARs not having i in the subset from the collection and then removing i from all subsets to effectively make their length $k - 1$ (steps 7-16).

Lemma 5.10. *Computational complexity of Algorithm* 5.5 *is* $O(N^3)$.

Proof. Let the time complexity of the algorithm is $T(N)$. All steps excluding the recursive call of the algorithm take $O(N^2)$ time as $|S| \leq N \times \binom{N}{N-1} = O(N^2)$. So, we can express $T(N)$

Algorithm 5.5: $N_D = \textbf{Decode\_FDGAS\_R}(N, k \equiv N-1, S, \&(da_1, \dots, da_N), I)$

Input:

- Number of POIs, N
- Degree of desired anonymity, $k = N-1$
- Set of already generated ARs, S
- Decoded attributes, $(da_1, \dots, da_N)$
- Set of POIs, $I = \{1, \dots, N\}$

Output:

- Number of decoded attributes, N_D

1. IF $S = \emptyset$ THEN
2. Set $a = \underset{\forall a \mid \exists j:(j:a) \in S}{\operatorname{argmax}} |\{ss|(ss:a) \in S\}|$
3. Set $S' = \{ss|(ss:a) \in S\}$
4. IF $|S'| = N$ THEN
5. Set $da_i = \infty$ for all $1 \le i \le N$
6. Set $N_D = 0$
7. ELSE IF $\exists i \in I: S' = PAS_i^{I,k}$ THEN
8. Set $da_i = a$
9. Set $S'' = \left\{(ss \setminus \{i\}: b) \middle| (ss:b) \in S \setminus S' \wedge ss \in PAS_i^{I,k}\right\}$
10. Set $N_D' = \textbf{Decode\_FDGAS\_R}(N-1, k-1, S'', \&(da_1, \dots, da_N), I \setminus \{i\})$
11. IF $N_D' = 0$ THEN
12. Set $N_D = 0$
13. ELSE
14. Set $N_D = 1 + N_D^{'}$
15. END IF
16. END IF
17. ELSE
18. Set $N_D = 0$
19. END IF

with the following recursive expression for the complete decodability case, which is also the extreme scenario:

$$T(N) = \begin{cases} O(1), & \text{if } N = 1; \\ T(N-1) + O(N^2), & \text{otherwise.} \end{cases}$$

By expanding the recursive expression we get

$$T(N) = T(N-1) + N^2 = T(N-2) + N^2 + (N-1)^2 = \cdots = \sum_{i=1}^{N} i^2 = O(N^3).$$ ■

In the next section, we discuss an important implementation issue regarding how the possible variation in attributes over time would be addressed.

5.4 Temporal Fluctuation of Attributes

Recalling from Section 4.5.54.4.1, considering the frequency of attribute change, an efficient statistical model can be developed to find the expected fraction of time, Γ_R by which ApS recovers the correct knowledge about all POI-attribute associations. This implies a time gap within which the ApS can provide dependable service. We can now analyse the transient effect of attribute change on the decodability while using DGAS. Since the change in attribute value (say, price change at a petrol pump) causes contradictions with previous reports from the corresponding POI, the decodability performance of the decoder degrades and will take some time to recover. Similar to what we did in the previous chapter for PGAS, we may empirically ascertain the expected length of Γ_R to remain preferably 90% of the full temporal window. Here, we present an analytical model to determine the expected value of Γ_R.

Consider a participatory sensing system of N POIs that change their respective attributes a_i's independently at exponentially distributed intervals with mean μ_i, $1 \leq i \leq N$. Effectively, the interval between changes of at least any two attributes is also exponentially distributed with mean $\hat{\mu} = 1/\sum 1/\mu_i$. Here, the attribute change interval, μ is the key parameter in the fluctuation modelling.

Let $n_a = n \times \omega$, number of participants observe these POIs, anyone at random, independently at exponentially distributed intervals with mean β_j, $1 \leq j \leq n_a$. For sufficiently large $\hat{\mu}$, we may safely assume that between changes of at least any two attributes j-th participant observes each POI on average $\hat{\mu}/N\beta_j$ times and collectively $\hat{\mu}\sum 1/\beta_j = (\sum 1/\beta_j) \,/\, (\sum 1/\mu_j)$ observations are made.

Let $m_{N,k}$ be the average number of reports needed to decode attributes of all N POIs using k-anonymity. So, the application server is expected to be knowledgeable of all attributes for the fraction of time,

$$\Gamma_R = 1 - \frac{m_{N,k} - 1}{\sum \frac{1}{\beta_j} / \sum \frac{1}{\mu_j}} = 1 - \frac{\sum \frac{1}{\mu_j}}{\sum \frac{1}{\beta_j}} \left(m_{N,k} - 1\right). \tag{5.2}$$

For a simpler case where all POIs are changing attributes in same interval and all the n_a participants are reporting in the same observation interval, i.e., if $\mu_1 = \cdots = \mu_N = \mu$ and $\beta_1 = \cdots = \beta_{n_a} = \beta$ then,

$$\Gamma_R = 1 - \frac{N\beta}{n_a \mu}\left(m_{N,k} - 1\right) \tag{5.3}$$

Here m_{N_k} is a function of (N, k), number of active participants n_a depends on the popularity of the scheme, while μ is governed by the POIs. The users' observation frequency is the only variable parameter to design a reliable service scheme such that the ApS has correct knowledge about all POI-attribute associations for a user-defined period of time. The higher the value of Γ_R, the ApS sustains complete and exact information for a greater fraction of time.

In Section 5.5, we shall present the simulation results to demonstrate that the higher the value of N, it requires a greater number of reports to maintain full decodability in 90% or more times. Besides, for less frequently changing attributes, it requires a fewer number of reports to maintain full decodability in 90% or more of the time.

5.5 Performance Evaluation

In this section, we present simulation results to establish the superiority of the decoding performance of DGAS compared to PGAS. The number of observations needed to achieve a certain level of data integrity indicates the performance of the respective approach. The simulation setup was the same as for PGAS in the previous chapter (presented in Section 4.5.1). To recall in brief, MNs report observations to AS in a random fashion that anonymizes those using respective algorithms and also the knowledge of previous reports. The results are generated for different degrees of anonymity, i.e., $k = N - 2$ to $N - 1$. All the results presented here are obtained by averaging 1000 simulation runs. Note that the decodability is computed here in a deterministic manner for all the schemes as opposed to the majority decoding based decoding computation performed in Chapter 4.

The performance of our proposed DGAS in achieving a certain level of data integrity and withstand attribute fluctuation, a comparison with its efficient variation FDGAS and also with the previously presented PGAS are presented here. Note that, in the previous chapter the performance of PGAS was measured in terms of decodability percentage, i.e., among N number of POIs which percentage of POIs was decoded. From now on, the performance of DGAS as well as comparison of that with its other variants are undertaken on the basis of full decodability percentage, i.e., the percentage of achieving N-decodability.

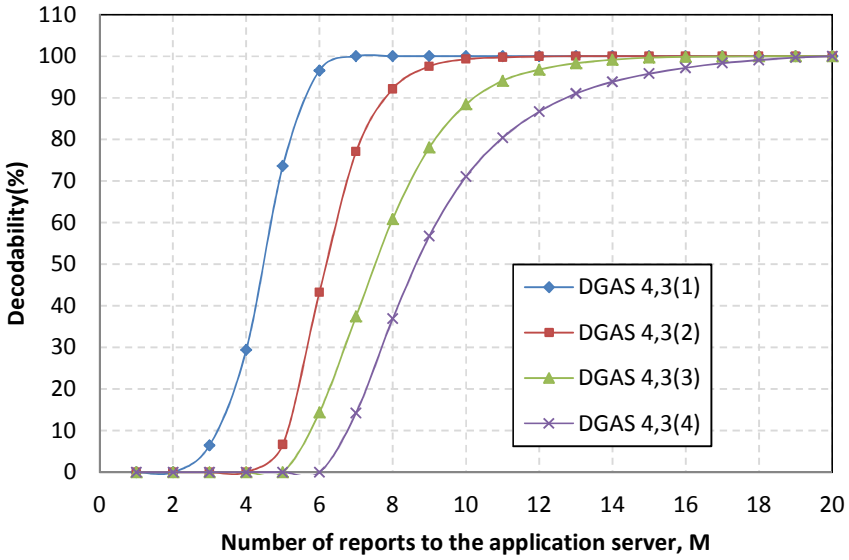


Figure 5.4: Data integrity trend when $N = 4$ and $k = 3$, for all possible D-decodabilities, where $D \in \{1,2,..,N\}$ generated from M number of observation reports.

5.5.1 Decodability Performance of DGAS

Figure 5.4 elaborately describes how D-decodability performance varies for different values of D, for DGAS with a fixed value of $N = 4$ and $k = 3$. Performance curve for each set up is identified as $N, k(D)$ where $D \in \{1,2,..,N\}$. The trends confirm that, a certain proportion of the simulation runs achieved D-decodability with fewer ORs when D is lowered. As the desired level of decodability, i.e., value of D increses, the value of required M also needs to be increased. To achieve 1-decodability i.e., to become confirm about one POI-attribute association among the total of four POIs where desired anonymity, $k = 3$, the number of required reports to ApS, $M = 7$. However, to arrive at full decodability the number of necessary reports, $M = 18$. This signifies that required value of M increases with desired level of D.

5.5.2 Fluctuation of Attributes

In this simulation, we presented different types of transient effects of POI attribute change. Here, the number of active participants $n_a = 200$. Figure 5.5 considers a generic case where, each POI changes attributes on average once per day. To maintain knowledge about POI-attribute association for 85% of time, the ApS needs 57 obseravations per POI per day for $N = 4$, whereas that required for $N = 5$ is 96, and $N = 6$ is 200. From another point of

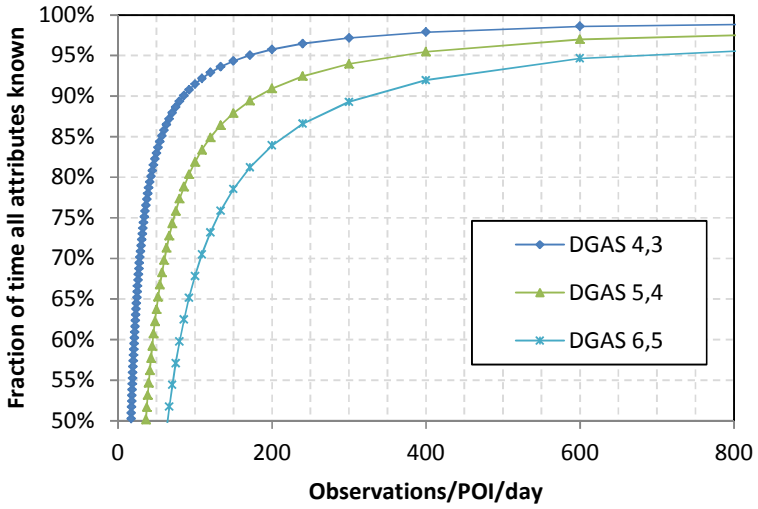


Figure 5.5: Effect of N on number of reports required to maintain knowledge particular fraction of time considering $N = \{4,5,6\}$ and $k = N - 1$.

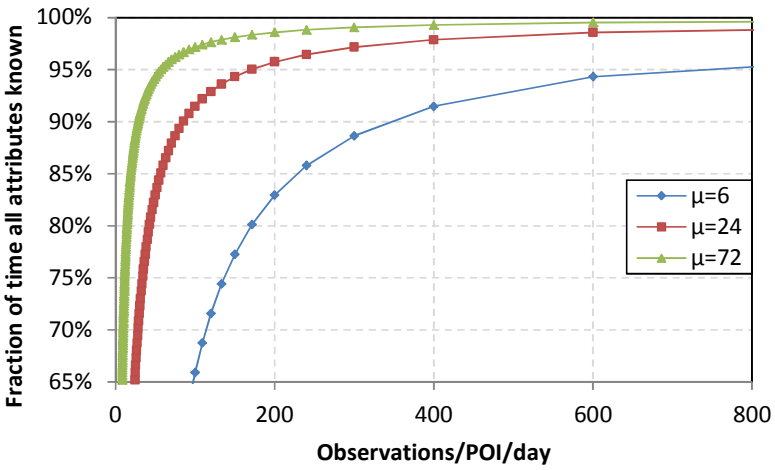


Figure 5.6: Effect of varying POI attribute change interval, μ on number of reports required to maintain knowledge particular fraction of time for $N = 4$ and $k = N - 1$.

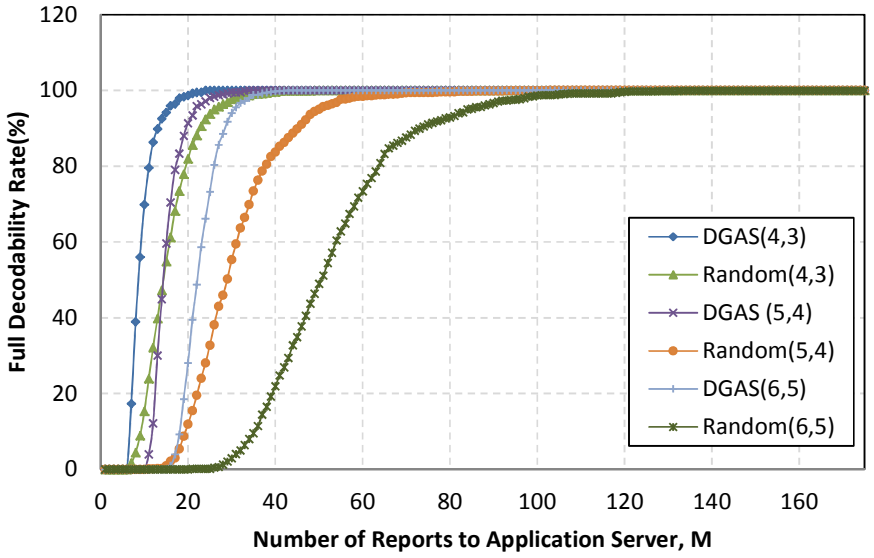


Figure 5.7: Data integrity trend when $N \in \{4,5,6\}$ and $k = N - 1$, for DGAS and Random scheme as compared from M number of observation reports.

view, when number of observation per POI per day is fixed to be 100 and $N = 5$, Figure 5.5 shows that for 82% fraction of time, our ApS is expected to be knowledgeable of the full decoded information. This is the requirement for highest possible privacy preservation, as in all cases we assume desired level of anonymity $k = N - 1$.

Then we consider varying the key fluctuation modelling parameter μ, for a fixed $N = 4$, and $k = 3$. On the one hand, for more frequently changing attributes, e.g., when POIs change attribute in every 6 hours, i.e., $\mu = 6$, Figure 5.6 shows that 300 observations per POI per day is required to retain 90% fraction of the time the ApS has knowledge about all attributes. On the other hand, when POIs change attribute less frequently e.g., in every three days, i.e., $\mu = 72$, only 29 observations per POI per day is sufficient to maintain the full decodability knowledge for 90% fraction of time.

5.5.3 Comparison between DGAS and PGAS

Figure 5.7 shows how the decodability performance of DGAS can be compared with its random variant for different values of N and $k = N - 1$. The performance curve for each set up is identified as (N, k) where $N \in \{4,5,6\}$. The figure shows that in all cases, a certain proportion of the simulation achieved 100% full decodability with significantly fewer observations when we used our proposed DGAS approach instead of the random one. As

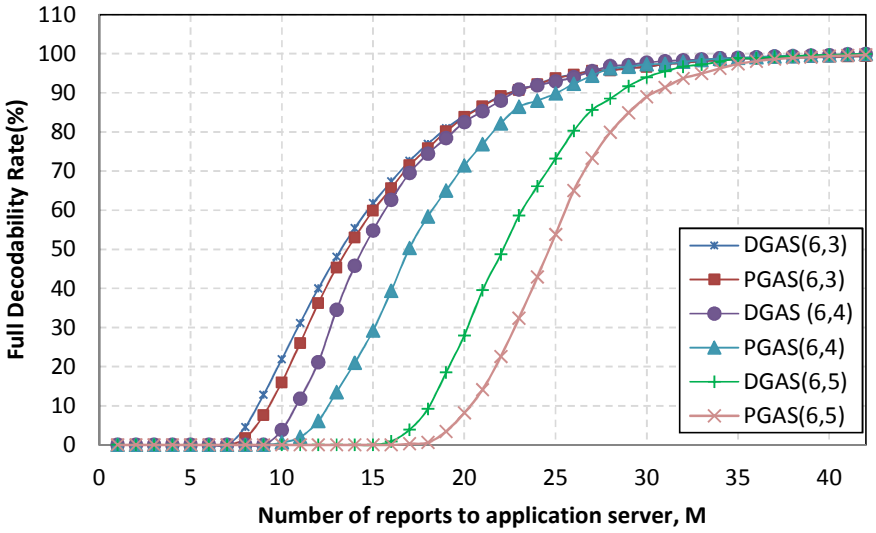


Figure 5.8: Data integrity trend when $N = 6$ and $k \in \{3,4,5\}$, for DGAS and PGAS as compared from M number of observation reports.

the number of POI increases, the gap between the decodability performance curves also increases. When $N = 4$, our proposed algorithm can reach 100% full decodability with reports $M = 24$, whereas for the same performance, the random scheme requires $M = 49$. This gap in required number of M to achieve the same performance increases from 25 to 133 using our proposed approach as compared to the random one, when number of POI is increased from 4 to 6.

Figure 5.8 presents the comparative performance of our proposed DGAS using trend lines with PGAS for the same values of N and k. Here, we consider, $N = 6$ and $k \in \{3,4,5\}$, to examine the effect of varying k as well. For all cases, a certain proportion of the simulation runs achieved 100% full decodability with almost same ORs for both DGAS and PGAS. However, prior to reaching 100% full decodability, DGAS is running over PGAS. Achieving up to 95% full decodability, DGAS requires a fewer number of observations compared to PGAS. For example, our proposed DGAS can reach 28% full decodability with $M = 20$, whereas PGAS can reach only 8.2% full decodability with same number of ORs, when $k = 5$. This gap by which DGAS is superior to PGAS, is increasing when operating with higher degree of anonymity, i.e., higher value of k.

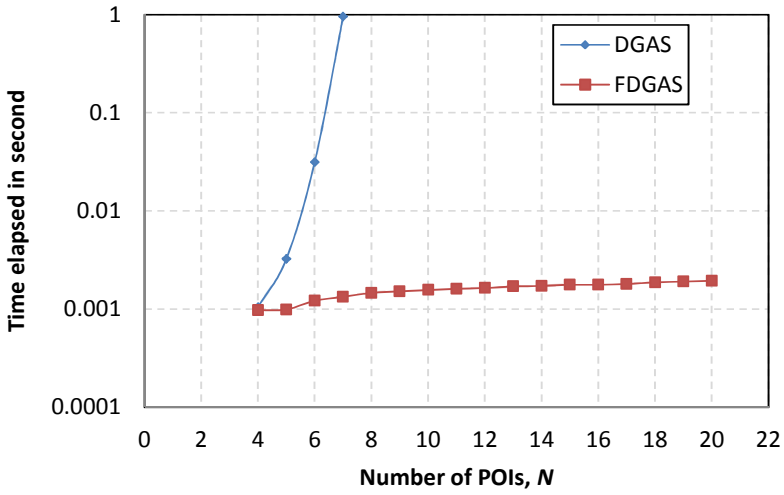


Figure 5.9: Time requirement for Encoding as compared between DGAS and FDGAS.

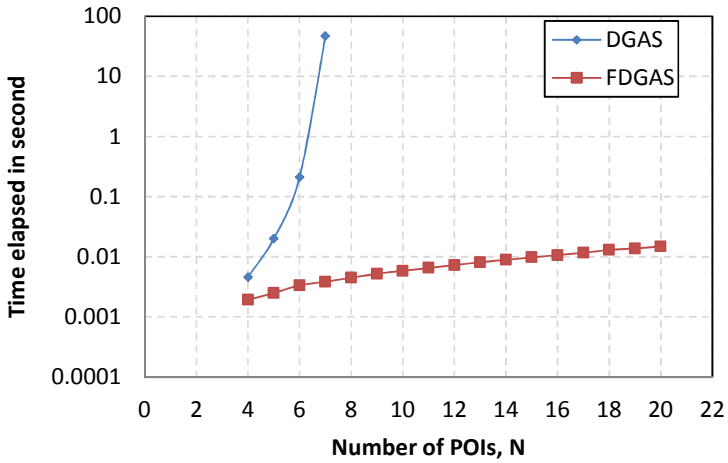


Figure 5.10: Time requirement for Decoding as compared between DGAS and FDGAS.

5.5.4 Comparison between DGAS and FDGAS

In some application scenarios, the privacy requirement might be the highest possible i.e., $k = N - 1$. For those cases we have developed a fast approach, FDGAS in Section 5.3.1 which is less complex both in computation and memory requirement. Figures 5.9 and 5.10 compare the time requirement of FDGAS as compared with DGAS in performing encoding and decoding respectively. The results clearly show that the complexity increases with the number of POIs. More importantly, FDGAS can compute at a much lower range of time compared to DGAS such that we needed to use a logarithmic graph to display them in the

same chart. Our original DGAS scheme can comfortably handle POIs in the range of 4 to 6, hardly 7 or 8. In contrast, if we used our FDGAS approach, time elapsed in computing is merely 0.76 milliseconds, even when N is as high as 20.

5.6 Conclusion

In this chapter, we have presented two more k-anonymization schemes using subset-coding which simultaneously protect the location privacy of the participating users and achieve the desired data quality at the target end. The deterministic approach based on cardinality reduction from a set of mapping possibilities between POIs and attributes was found to be superior to the probabilistic approach presented in the previous chapter in achieving data quality from fewer observations. The faster variation of this approach was also analysed empirically. In the next chapter, we are going to apply another subset-coding based anonymization approach for a completely different application scenario, electronic voting.

Chapter 6

6 Subset Coding Based E-Voting

In the previous two chapters we have presented subset-coding based *k*-anonymization techniques to protect the privacy of users in PSS. In this chapter, we use the same subset-coding as a basis of privacy preserving electronic voting that is simultaneously trustworthy to the voters who can verify that their votes are properly counted to produce an election outcome. The proposed voting system resorts to joint de-anonymization of the votes for counting ensuring that it is difficult to manipulate votes by any entity concerned without being detected.

The organization of the rest of this chapter is as follows. Section 6.1 constructs the introductory baseline of this chapter. We briefly discuss the idea of *k*-anonymity based voting in Section 6.2. In Section 6.3, we present the voting protocol in details. We present the post-voting procedures, including tallying and auditing, in Section 6.4. In Section 6.5, we analyze various threats to the proposed voting scheme. Some optimization issues related to our proposed scheme are discussed in Section 6.6. Finally, we conclude the chapter in Section 6.7.

6.1 Introduction

Democracy, the most wide-spread political system in the world, relies a great deal on free and impartial election. It is believed that [125] if voters can trust an election, their participation would be spontaneous and they will put in their best effort to choose the right candidate. A trustworthy voting system has to be verifiable by the voters individually and also globally by some responsible parties. However, the indispensable requirements of voting such as confidentiality and privacy of the voters also must be respected.

Among all the requirements discussed in detail in Section 2.2, the following are of critical importance to a trustworthy voting system.

Fairness – all the eligible voters should have equal weight to influence the outcome of an election and nobody should be able to influence the outcome of an election more than with her own vote.

Privacy – it should be impossible for anyone to know how an individual voter voted. It should also be impossible for a voter to prove to anyone how he voted. The former requirement also ensures *coerce-resistance* while the latter prevents *vote trading*.

Integrity – it should be impossible for anyone, including the election officials, to tamper with the voting results. The measures to ensure the integrity of a voting system is verifiability, i.e., it should be possible to verify that each and only the authorized votes are counted. In other words, E2E verifiability has to be maintained.

To achieve integrity, in addition to verifiability, a voting system needs to ensure authenticity, i.e., a voter or a group of voters must not be able to prove a false claim about the election result. To achieve fairness, a voting system should ensure that all eligible voters are allowed to vote only once and no eligible voter is denied her voting right. While verifying eligibility, the voting system should also maintain anonymity [22]. Therefore, the

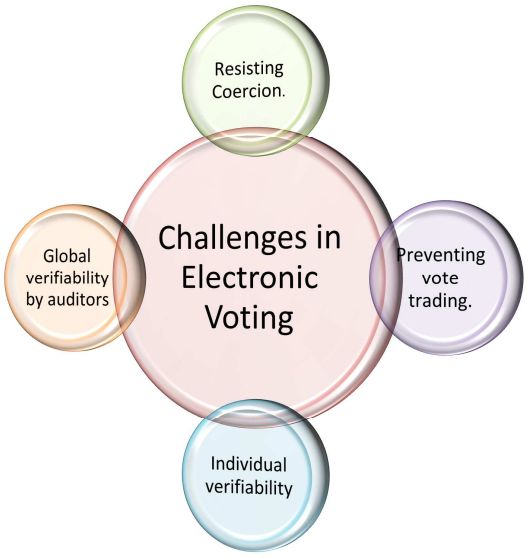


Figure 6.1: The challenges of an EVS.

traditional method where the eligibility is verified in front of the voting booth can be used to ensure fairness. Thus, the major objective of an electronic voting system is to ensure both privacy and integrity. This is a difficult problem since privacy mandates that there would be no link between the vote and the voter, whereas integrity or trustworthiness requires verifiability which needs to preserve association between a vote and its originator. Verifiability is desired individually by the voters as well as globally by some third party auditors. All these simultaneous requirements are summarised in Figure 6.1.

Electronic voting, having the privilege of fast and flawless counting, has over the years been explored to meet all these demands. Some schemes have tried to combine the simplicity of the traditional system with some security mechanisms using electronic machines. However, to provide trustworthiness of the system, these techniques either assume unrealistic mechanisms, such as use of special inks, or demand unrealistic voter capability, such as performing modular arithmetic or trusted entities such as Electronic Voting Machine (EVM), printer, scanner, etc [22]-[28]. In reviewing the literature, there is yet to be any universally acceptable electronic or electronic device-aided voting system.

In this chapter, we present a voting scheme that ensures minimal trust assumption of the machines to achieve both privacy and E2E verifiability of the voters. To ensure the voter privacy we use a k-anonymization scheme based on our subset-coding technique introduced in Chapter 3. The idea of our approach is to anonymize a vote with $k-1$ other candidates. The problem of voting is quite different from the natural context where subset-coding is applicable. We have mapped the voting problem to this context by considering a group of candidates who would be presented to a number of voters similar to a collection of observers visit who a POI. Our hypothesis is that the individual anonymized votes can be jointly decoded, provided a sufficient number of votes are cast.

The feasibility of the technique relies on the efficient joint de-anonymization property of subset-coding technique to obtain the election outcome. To ensure the trustworthiness of the electronic entities (hardware and communications) involved in the voting system, each of the entities' behaviour is kept under check and balance by other entities in the system. Essentially, this ensures that any manipulation attempt by any of the entities will be detected by other components of the system with high probability. The proposed voting scheme also uses standard security/privacy measures such as cryptographic hashing while storing and/or transmitting the casted votes to the central authority and/or Bulletin Board

(BB). We also incorporate the idea of a floating receipt as proposed by [27] to protect against vote trading or coercion. To ensure receipt-authenticity, we propose using both a digitally signed paper and on-the-spot physical validation by a polling officer.

Specifically, the proposed *k*-anonymity based voting scheme has the following comparative advantages over other existing systems:

- The proposed scheme ensures trustworthiness of the whole system by distributing the responsibilities among the entities. The behaviour of each of the entities is verified by one or more other entities of the system. Therefore, this does not require assuming expensive and/or infeasible hardware or voter capability to achieve trustworthiness of the system.
- The existing schemes use pseudorandom permutation of the candidates to protect the privacy of each of the votes. Our approach instead checks the consistency of the votes during the joint de-anonymization. As we show in Section 6.5.1, even a negligible number of vote manipulations will not remain undetected in this system due to this inherent consistency preservation property. This allows our proposed system to adopt a simple hardware model.
- Some existing systems such as [25] protect voter privacy by allowing partial information about the vote in a receipt and eventually fail to ensure complete verifiability. In contrast, the proposed system provides full voting information in the receipt and, thus, ensures E2E verifiability.

In the following sections, we present our voting scheme.

6.2 Preliminaries

In this section, we discuss the idea of how subset-coding based anonymization scheme is designed for a completely different application scenario. Recall from Chapter 3 that subset-coding is applicable to any multiple observation of individual instance scenario. We have mapped the voting problem to this context by considering a group of candidate a single object on which a collection of voters would cast their preference. To present the preliminary concept of our approach we use an example as follows.

Let there be $N = 4$ candidates A, B, C, and D and we want to achieve $k = 2$ anonymity. Let $[A\ B\ C\ D]$ be the canonical ordering of the candidates. Clearly we can have $4! = 24$

possible ordering for these candidates. Now in the voting scheme, a voter is randomly provided with a candidate-order out of these 24 possible orderings. Let a voter is assigned the candidate-order $[B\ C\ A\ D]$ and her favorite candidate is C. Now she needs to construct a subset of $\{A\ B\ C\ D\}$ containing exactly $k = 2$ candidates which should include the candidate C. Thus, her possible choices are$\{A, C\}$,$\{B, C\}$, and $\{C, D\}$. Assume that her choice is$\{A, C\}$. Then, her completed vote should be recorded as$\{A, C\}: 2$, since the index of C in the candidate-ordering $[B\ C\ A\ D]$ is 2.

To count the votes, the central election authority, as well as a verifier, must be able to decode the votes. Clearly knowing the candidate-ordering used in a particular vote, allows anyone to decode the vote. For example, given that the candidate-order $[B\ C\ A\ D]$ was used while casting the vote $\{A, C\}: 2$, one can easily decode that the vote is for candidate C. However, the major strength of the scheme is that the election authority or verifier can decode the votes without knowing the candidate-ordering used to cast the vote as long as one has a reasonable number of anonymized votes. In the instance in the current example, let a group of voters voted using the candidate-ordering, $[B\ C\ A\ D]$ as $\{A, C\}: 2$, $\{B, C\}: 2$, $\{B, C\}: 1$, $\{A, B\}: 3$, $\{A, D\}: 4$. This can be verified that among the possible 24 candidate-orderings, only $[B\ C\ A\ D]$ is consistent with all these votes. We can now calculate the number of votes for candidates A, B, C, and D as 1, 1, 2, and 1, respectively.

Clearly the brute-force approach to de-anonymize the votes considering all the $N!$ possible candidate-orders is not computationally feasible. We also need to ensure complete de-anonymization even in case of an insufficient number of votes to achieve full de-anonymization (to be defined by Definition 6.4). In Section 6.4, we describe how the de-anonymization can be done efficiently. To address the problem of insufficient number of votes for complete de-anonymization we propose the idea of dummy votes in Section 6.4.1.

6.3 Voting Protocol

In this section we present the whole voting process including pre-voting steps and the functionalities after the arrival of each voter in the booth.

The proposed voting system involves the following hardware entities: (i) random number generator (RNG); (ii) electronic voting machine; (iii) receipt exchange box; (iv) public Bulletin Board (BB); and (v) central server. Human entities involved with the system are: (i) voters; (ii) multiparty polling agents; (iii) vote counting authority; and (iv) auditors. Before describing the voting protocol in detail, we first define relevant notions.

Definition 6.1 (Candidate-order) *Let* $[C_1\ C_2 \cdots C_N]$ *be the canonical ordering of* N *candidates. Then an indexed candidate-order*

$$[C_{i_1} C_{i_2} \cdots C_{i_N}] \leftarrow Perm\ (s, [C_1\ C_2 \cdots\ C_N]\ ), \tag{6.1}$$

accepts s *as the index and the canonical candidate-order* $[C_1\ C_2 \cdots\ C_N]$ *and outputs a permutation* $[C_{i_1} C_{i_2} \cdots C_{i_N}]$ *of the canonical candidate-order.*

Definition 6.2 (Permutation group) *Given a set of* $L \leq N!$ *distinct indices* $\{s_1\ s_2 \cdots\ s_L\}$, *the permutation group* G *is defined as the set of candidate-orders corresponding to indices* $s_j, j = 1, \cdots, L$, *i.e.*

$$G\ =\ \{Perm\ (s_j, [C_1\ C_2 \cdots\ C_N])\}, j = 1, \cdots, L. \tag{6.2}$$

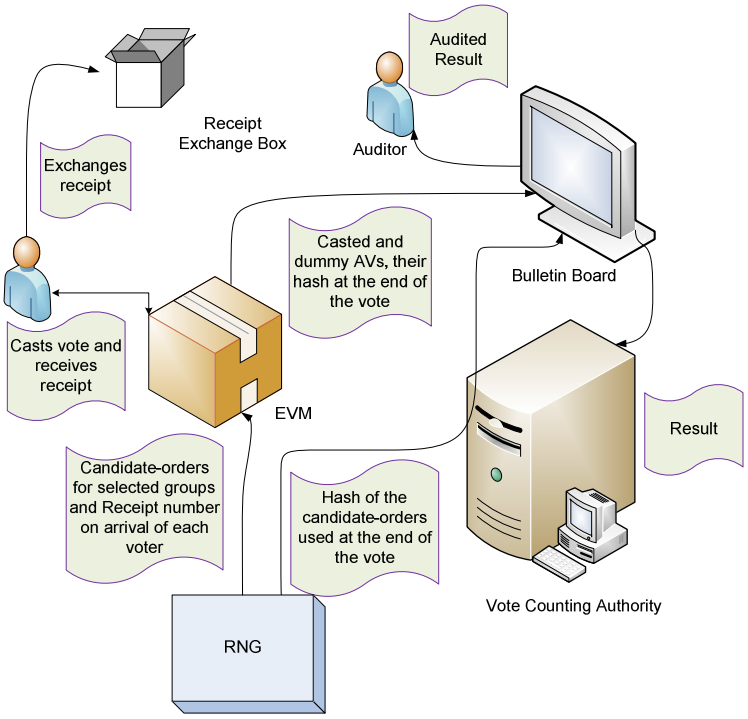


Figure 6.2: Conceptual diagram of the proposed electronic voting process.

The whole voting process is depicted in Figure 6.2 . In the following section, we describe the pre-voting tasks and the procedures that an eligible voter will experience in the course of casting a vote.

6.3.1 Receipt Number and Candidate-order Generation

Given the number of candidates N, the RNG chooses the permutation group size L at the beginning of the election. The RNG then generates L indices and the corresponding L candidate-orders are then provided to the EVM. The RNG also computes a hash value of the permutation group and posts it to the BB after the completion of the vote. This hash value enables the public to verify that the EVM used the correct permutation group. To ensure the trustworthiness of the RNG, its behavior will be controlled by the seed that is provided by the multiparty pooling agents at the beginning of the election. Later in this chapter, we show that a compromised RNG has very limited capability to manipulate the election in favor of a particular candidate.

On the arrival of an eligible voter, the RNG generates a receipt containing a unique randomly generated number r of length n which is also supplied to the EVM. The EVM uses a one way function $f(\cdot)$ that maps r to $\{1, \cdots, L\}$ to choose a candidate-order from the permutation group supplied by the RNG at the beginning of the election. Let r_i denotes the i-th digit of r. Then in the proposed system we assume the following structure for the function $f(\cdot)$,

$$f(r) = (a_0 r_0 + a_1 r_1 + \cdots + a_n r_n) \bmod L, \quad \text{where } a_i \in \{0,1\}, i = 1, \ldots, n. \tag{6.3}$$

The EVM randomly selects the values of a_i's that determine the behavior of $f(\cdot)$. Indeed, due to the use of this function, an untrusted RNG cannot manipulate the election in favor of a particular candidate. To ensure that the EVM is consistent with the values of a_i's throughout the election, it is a requirement that the EVM post the values of a_i's to the BB at the end of the voting.

To prevent any collusion between the RNG and EVM, we assume that the communication link between RNG and EVM is one way, i.e, from RNG to EVM only. This one way communication can be achieved as follows: there will be no electronic communication between the RNG and EVM; the output generated by the RNG printed on the receipt will be fed to the EVM by the voter herself, and to facilitate the reading of the

receipt number by the EVM, the RNG will print the number along with its barcode in the receipt.

After receiving a receipt from a voter, the EVM reads the number r and chooses the candidate order corresponding to $f(r)$. Then the EVM simultaneously displays the receipt number r and the candidate-order corresponding to $f(r)$ to the voter. At this point, there is a risk of manipulation by the EVM if it does not use the original receipt number r. Therefore, the handling of the receipt by the EVM would be done using physically transparent hardware so that the voter gets the original receipt at the end of her vote. Another possible fraudulent attempt by a compromised EVM may be to choose and display a different candidate-order other than that corresponding to $f(r)$. We term this as group-interchange attack. Since, at the end of the voting, the hash value of the given permutation group and the values of a_i's will be posted to the BB by the RNG and EVM respectively, the EVM cannot use any other number without being detected due to the inherent strength of joint de-anonymization. We analyse this attack and also the detection probability in Section 6.5.1.

The next step is to cast the vote in k-anonymized form. We discuss this step below.

6.3.2 Casting Vote

In the proposed system, the voters cast their votes using the k-anonymization technique discussed here. Let us term the k-anonymized form of a vote as Anonymized Vote (AV),

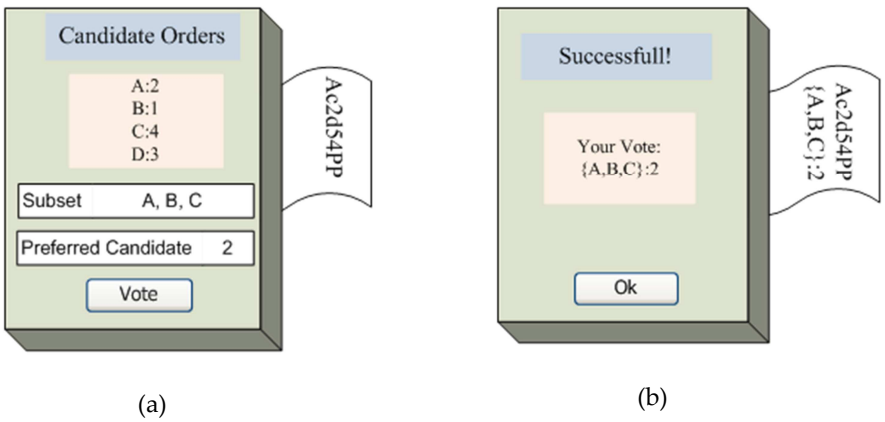


Figure 6.3: For a set of 4 candidates $\{A, B, C, D\}$,(a) the voter inputs her preference in 3-anonymized way following the displayed candidate-order and half-printed receipt with receipt-number only and (b) the EVM shows the recorded vote while printing the full receipt.

Table 6.1: Possible Anonymized Votes for the Candidate-order $[C\ A\ D\ B]$

Vote for A	Vote for B	Vote for C	Vote for D
$\{A, B, C\}$: 2	$\{A, B, C\}$: 4	$\{A, B, C\}$: 1	$\{A, B, D\}$: 3
$\{A, C, D\}$: 2	$\{A, B, D\}$: 4	$\{A, C, D\}$: 1	$\{A, C, D\}$: 3
$\{A, B, D\}$: 2	$\{B, C, D\}$: 4	$\{B, C, D\}$: 1	$\{B, C, D\}$: 3

defined as follows.

Definition 6.3 (Anonymized Vote): *Given a candidate-order* $\left[C_{i_1} C_{i_2} \dots C_{i_N}\right]$*, an Anonymized Vote (AV) for casting in favor of the candidate* $C_{i_j}, 1 < j \leq N$*, is expressed as*

$$\left\{C_{i_j}\right\} \cup \left\{C_{j_1}\ C_{j_2} \cdots C_{j_{k-1}}\right\}: i_j, \tag{6.4}$$

where $\{j_1, j_2, \dots, j_{k-1}\} \subset \{1, \dots, N\}/\{i_j\}$ *is selected by the voter to anonymize her vote with other candidates.*

After the completion of a vote casting, the EVM appends the anonymized vote $\{C_{i_j}\} \cup \{C_{j_1}\ C_{j_2} \dots C_{j_{k-1}}\}: i_j$ below the receipt number and gives the complete receipt to the voter. Since the receipt was already printed with r and the hash value of the permutation group and the values of $a_i{}'s$ defining $f(\cdot)$ will be made available to the BB, the EVM cannot manipulate the candidate-order. The vote casting protocol is depicted in Figure. 6.3 for a typical scenario with four candidates.

Example: Let $A, B, C,$ and D be four candidates and $[A\ B\ C\ D]$ be their canonical order. For a candidate-order $[C\ A\ D\ B]$, Table 6.1 shows a list of the possible AVs to maintain $k = 3$ anonymity where the voter wishes to vote for $A, B, C,$ or D. If she chooses to vote for B, she can choose any of the three possible AVs: $\{A, B, C\}$: 4, $\{A, B, D\}$: 4, or $\{B, C, D\}$: 4. To prevent potential manipulation of the EVM, the voter is given the control to choose any of the possible subsets. If the voter inputs an inconsistent AV, EVM will detect it and ask the voter to vote again. For example, $\{A, B, C\}$: 3 is an inconsistent AV for the candidate-group presented above.

The EVM stores all the AVs in its memory to send the votes to the central election authority and to the BB at the end of the voting session. To enable de-anonymization of the anonymized votes, the stored AVs also contain the identification of its originating EVM.

The receipt collected after each vote preserves an association between a vote and the voter which makes the privacy of the voter vulnerable. Our target is to use the receipt for individual verification, however, without keeping any link to its originator. This is achieved by swapping receipts, as discussed below.

6.3.3 Floating the Receipt

After collecting the receipt from the EVM, the voter comes to the polling officer to validate her receipt with a signature and seal. Together with the digital signature, this manual validation prevents fake-receipts[1]. After validation, the polling officer puts it into a transparent receipt exchange box and guides the voter to randomly collect another receipt from the exchange box. This mechanism is termed as *floating receipts,* and as proposed in [27] it decouples the link between a voter and the receipt she carries. Distributed or individual verification is achieved by expecting that a large number of voters will verify the collected receipts in the BB.

It may be argued that the voters have more motivation to check their own receipt, rather than checking an unknown voter's vote. However, the joint de-anonymization ensures that all the original AVs will jointly conform to a particular candidate-order and any fraudulent attempt will get noticed by bringing inconsistency to the de-anonymization process. In the proposed scheme, if a voter checks and verifies a floating receipt from the BB, she can be assured that her vote is also recorded-as-cast. The strength of joint de-anonymization is further elaborated in Section 6.4.

Next, we discuss the post-voting procedures run at the central authority and also by third-part auditors to count the votes and ensure the global verifiability by detecting potential manipulation attempts by the EVMs.

6.4 Post Voting Procedures

At the end of the voting session, the polling agents count the number of voters and publish the total number of votes cast. This count will be matched against that provided by EVM. After this, using a cryptographic hash function, the EVM computes a hash value or message digest based on all the information stored in the ROM which includes AVs and the values of $a_i's$ defining $f(\cdot)$. The message digest is handed over to the polling agents of the different

[1] However, if it can be extremely difficult to fake the receipt, this step can be skipped.

parties before transferring the content of the ROM to the central election authority and/or to the BB. The total number of votes in each EVM is counted by the polling agents and is separately provided to the central vote counting authority and the BB. The RNG also computes the hash value of the permutation order given to the EVM at the beginning of the voting session and is handed over to the polling agents. A major strength of the proposed joint de-anonymization technique is that all the AVs of a permutation group are consistent with one and only one candidate-order. This consistency property ensures that any fraudulent attempt by the EVM to change candidate-orders would be detected as a contradiction during vote counting.

In the case of physically transferring the ROM, the transparency can be ensured by physical monitoring of different polling agents. In the case of electronic transfer, encryption and standard secure channels can be used. If a channel intruder makes any change to the AVs, the message digest will mismatch and, therefore, the change will be detected. Note that it is computationally infeasible to modify a message without changing the message digest and it is also computationally infeasible to find two different messages with the same message digest. This property also ensures that if any change is made to any data sent from an EVM, it can be detected by re-computing the message digest from the same data presented in the BB.

The central authority processes the collected AVs from all the EVMs in the central server and presents them in the BB so that an individual can match her receipt against the one in the BB. The message digest is also presented and verified by the polling agents. The authority de-anonymizes these AVs using the technique discussed below, counts votes for each candidate, and then declares the vote result. Finally, the third party audit team plays the role of an external group that checks and verifies the election result to confirm the universal verifiability property using the same de-anonymization technique from AVs presented in the BB. They also compute a message digest from the raw data and match it against the one presented in the BB.

6.4.1 Tallying

Each AV needs to be de-anonymized for tallying. We present here the joint de-anonymization technique that uniquely associates each vote to the candidate that was actually cast, provided that a sufficient number of anonymized votes have been cast for each group. In the following, we formally define the notion of full de-anonymization.

Table 6.2: Conforming Tuples of Gradually Counted AV Set for Candidates $[A\ B\ C\ D]$

AVs	{A, B, C}: 1	{A, B, C}: 1 {A, C, D}: 3	{A, B, C}: 1 {A, C, D}: 3 $\{A, B, D\}$: 1	{A, B, C}: 1 {A, C, D}: 3 $\{A, B, D\}$: 1 {A, B, C}: 1 $\{B, C, D\}$: 4	{A, B, C}: 1 {A, C, D}: 3 $\{A, B, D\}$: 1 {A, B, C}: 1 $\{B, C, D\}$: 4 {A, B, C}: 2
Conforming tuples	(1,2,3,4) (1,2,4,3) (1,3,2,4) (1,3,4,2) (1,4,2,3) (1,4,3,2) (2,1,4,3) (2,1,3,4) (2,3,1,4) (2,4,1,3) (3,1,2,4) (3,1,4,2) (3,2,1,4) (3,4,1,2) (4,1,2,3) (4,1,3,2) (4,2,1,3) (4,3,1,2)	(1,2,3,4) (1,2,4,3) (1,4,2,3) (1,4,3,2) (2,1,4,3) (2,1,3,4) (2,4,1,3) (3,1,2,4) (3,1,4,2) (3,2,1,4) (3,4,1,2) (4,1,2,3) (4,1,3,2) (4,2,1,3)	(1,2,3,4) (1,2,4,3) (1,4,2,3) (1,4,3,2) (2,1,4,3) (2,1,3,4) (3,1,2,4) (3,1,4,2) (4,1,2,3) (4,1,3,2)	(1,2,3,4) (1,2,4,3) (1,4,2,3) (1,4,3,2) (2,1,4,3) (2,1,3,4) (3,1,2,4)	(1,2,3,4) (1,2,4,3) (1,4,2,3) (2,1,4,3) (2,1,3,4) (3,1,2,4)

Definition 6.4 (Full De-anonymization): *For an N-candidate election, a particular outcome of an anonymization scheme satisfies full de-anonymization iff all N candidates can be associated to their correct candidate-order.*

From the receipt number of an AV, one can easily compute the associated group number given the values of a_i's. However, the candidate-order corresponding to a group number is not directly available. Now given all the AVs belonging to a particular group, one can determine its candidate-order using the joint de-anonymization method (as presented below).

Let us explain the de-anonymization method with an example. In Table 6.2, we present the de-anonymization method for $N = 4, k = 3$ and a candidate-order $[A\ B\ C\ D]$. The algorithm starts with all possible mappings between the candidates and their orders. Each AV rules out some of these mapping possibilities and when only one possibility is left, full de-anonymization is achieved. In Table 6.2, the conforming tuples are shown each time an AV is taken into account. Note that the number of tuples gradually decreases with the

number of AVs. When a new AV is taken into consideration, some of the existing tuples conforming to previously considered AVs, contradict with the new AV. In the end, the remaining tuple/s may or may not lead to full de-anonymization.

Since the individual voters select the k candidates in the AV in an unguided manner, a group of votes may fall short of full de-anonymization. In this case the central authority would not be able to count votes from the received AVs. To overcome this problem, we propose to use dummy votes that will ensure unique and correct decoding. The dummy votes are candidate-group specific and easily distinguishable from the original ones while tallying. After the voting session, the EVM determines the minimal set of dummy AVs that will bring the group to full de-anonymization. Thus, the EVM also needs to send the set of dummy AVs, along with the original AVs, to the central authority. Now, we describe the procedure to compute a set of dummy AVs.

Let $S(N,k)$ be a set of AVs that yields full de-anonymization. Let $S_{\min}(N,k)$ be a set with a minimal number of AVs that achieves full de-anonymization. Now a $S_{\min}(N,k)$ can be constructed using the following technique.

Following the joint de-anonymization technique presented in Table 6.2, one AV eliminates $(N-k)$ possible candidate-orders. Again, to associate a candidate with her correct order, $(N-1)$ other possibilities need to be removed. Continuing with the candidate-order$[A\ B\ C\ D]$, the three AVs $\{A,B,C\}:1$, $\{A,B,D\}:1$, and $\{A,C,D\}:1$ suggest that D, C, and B cannot have the order 1. So, they jointly decide that the candidate-order of A is 1. To decode the next candidate, say B, we take into account the decoded order of A in the AVs. Accordingly, the AVs $\{A,B,C\}:2$ and $\{A,B,D\}:2$ will define that the candidate-order of B is 2. This process will continue until $k-1$ candidates are mapped to their respective orders. The remaining candidates may be mapped using only one AV by including any $k-1$ from the already decoded candidates. Here, to decode C, we use already decoded orders of A and B and find the AV $\{A,B,C\}:3$. For D, any two candidates from $\{A,B,C\}$ can be included. So, any of $\{A,B,D\}:4$ or $\{A,C,D\}:4$ or $\{B,C,D\}:4$ will serve the purpose, i.e., $\{-,-,D\}$.

According to this strategy, it is evident that the first candidate in a group can be mapped to its order using $\left\lceil\frac{N-1}{N-k}\right\rceil$ AVs. Similarly, for the second candidate it takes $\left\lceil\frac{N-2}{N-k}\right\rceil$ AVs and so on for the first $k-1$ candidates. For the remaining $N-k+1$ candidates, each requires only one AV. Hence we have the following proposition.

Proposition 6.5. $|S_{\min}(N,k)| = \sum_{i=1}^{k-1} \left\lceil \frac{N-i}{N-k} \right\rceil + (N-k+1)$. ■

At the end of voting session, EVM has to check if at least one $S_{\min}(N,k)$ can be constructed for each of the groups. If this is achieved, no dummy vote is required for that group. Otherwise, the EVM has to generate a minimal set of dummy AVs for the group, which together with the actual AVs, construct an $S_{\min}(N,k)$ for that group. For example, the final set of AVs shown in Table 6.2 need to include three AVs, i.e., $\{A,C,D\}:1$, $\{A,B,C\}:3$, and $\{A,B,D\}:2$ to achieve full de-anonymization. In other words, these three dummy votes along with the original AVs $\{A,B,C\}:1$, $\{A,B,D\}:1$, $\{B,C,D\}:4$, and $\{A,B,C\}:2$ would construct an $S_{\min}(N,k)$.

A simple way to generate the set of dummy AVs is to pre-decide a particular $S_{\min}(N,k)$ to be constructed and check if each of its AVs is present in the casted votes. If any AV from the target $S_{\min}(N,k)$ is not found, it has to be appended as a dummy vote.

6.4.2 Auditing

Once the voting is complete and the result is announced, the third party audit team plays the role of the universal verifier. They collect all the AVs from the BB and all the message digests from the EVMs. Then, the message digest from the AVs is computed and matched against the one published in the BB. They cross check the total vote counts as maintained by an EVM with the number of total votes casted. Finally, the audit team verifies and confirms the election result to uphold the universal verifiability property of the proposed scheme. For this purpose, they use the same de-anonymization technique as used by the central authority to get the final result from the AVs presented in the BB.

6.5 Threat Analysis of the Proposed System

First, we classify the possible threats into three types: *threats on verifiability, threat by manipulating dummy AVs, and threat on privacy.* We categorize the following threats into the class of threat on verifiability: (1) manipulation of the votes at the EVM, (2) alternation of the votes during transmission from EVM to central election authority and the BB, and (3) changing the votes at central election authority. The other threat is that the EVM can generate dummy AVs in such a way that the AVs become consistent with a different candidate-order other than the one displayed to the voter. Under the class of threat on privacy, we consider the threats that may breach the voter privacy and results in coercion

and/or vote trading that will indirectly affect the election. In the following section, we discuss how all these threats are accounted for in our proposed system.

6.5.1 Mitigation of Threat on Verifiability

The possibility of vote alteration is mitigated due to the use of cryptographic hashing since the hashing is done just after the voting session and published to the polling agents. External auditors/verifiers will compute the same hash for the votes counted at the central authority for each EVM and will publish these. The polling agents who have received the hash of corresponding EVMs will match this against the ones published at the BB and any mismatch here would be reported.

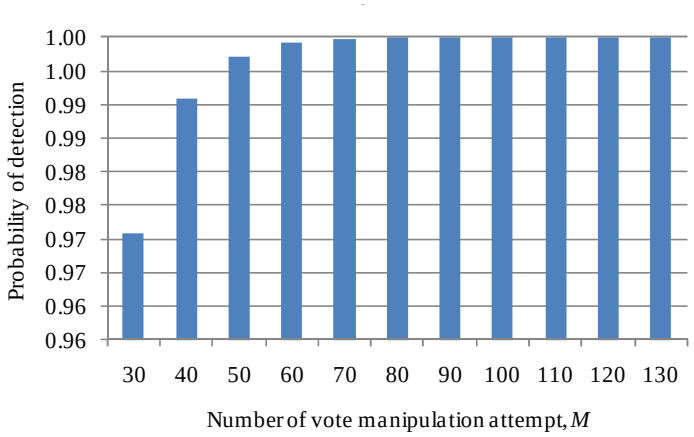
Individuals would find the receipt they carry in the BB and match the content. This would ensure that the vote is cast-as-intended and recorded-as-casted. The auditors compute the result as well to ensure all the votes are counted-as-recorded.

Now, the most likely chance of vote manipulation is in EVM and before the hash is published. As discussed in Section 6.3, when a voter comes to vote, the receipt number is printed by the RNG and handled by the EVM in a transparent manner. The voter inputs an AV with her preference and the full receipt comes out appending this AV. The only way an EVM may alter a vote is by presenting a candidate-order different from what is computed from the original receipt number r using $f(\cdot)$. Referring to two candidate-orders $a \equiv [A\ B\ C\ D]$ and $b \equiv [D\ B\ C\ A]$, let A be a popular candidate and EVM is compromised to interchange the votes between A and a manipulator candidate D. Then while $f(r)$ denotes candidate order b, the EVM would show candidate-order a to the voter. Consequently, a voter with a preference for A would choose AV as $\{A, B, D\}$: 1 or $\{A, C, D\}$: 1 and it would be counted as a vote in favour of D. Considering that the other candidates would not influence the interest of D, this is quite a plausible option for manipulation in favor of D. However, since the voter selects $k-1$ candidates for anonymization out of k options with an equal likelihood for everyone, there is a probability that the voter would not include the illicit candidate at all in the selected subset. In this example, $\{A, B, C\}$: 1 is that AV. This is clearly a contradiction for group b as neither of A, B, or C has candidate-order 1. It will be detected while counting votes and the corresponding EVM will be identified as compromised.

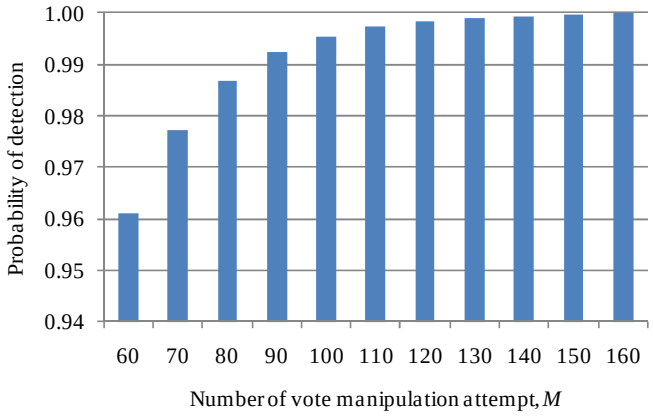
Let this potential attack on the integrity of our proposed electronic voting system by EVM be termed as a group-interchange attack. We may estimate the probability of detection

of a group-interchange attack by the inherent consistency property of the proposed joint de-anonymization technique. For N candidates, to achieve k-anonymity including the actual person to vote in the subset, the total number of possible subsets is $\varphi_1 = \binom{N-1}{k-1}$. Among these, $\varphi_2 = \binom{N-2}{k-2}$ subsets would include the manipulator candidate (D in the example above). Therefore, the probability P_U that group-interchange attack is undetected (that is equal to the probability of selecting the manipulator candidate in the subset) for a single vote be

$$P_U = \frac{\varphi_2}{\varphi_1}. \tag{6.5}$$



(a)



(b)

Figure 6.4: Probability of detection of group-interchange attempt by subset-coding technique for different number of votes attempted to be manipulated (M) when number of candidates (N) is (a) 10 and (b) 20.

If M votes are attempted to be changed, then for all of these the probability P_M of failure to detect any of these attempts is

$$P_M = (P_U)^M. \tag{6.6}$$

Then the successful detection probability P_D of M candidate-order interchange attempts by our technique is

$$P_D = 1 - P_U. \tag{6.7}$$

Since M is a reasonably high value for even a small scale election, P_D is very high and a compromised EVM will thus be easily identified. Figure 6.4 presents a numerical analysis of P_D for different number of votes attempted to be manipulated (M) for a permutation group. We have considered typical election scenarios to investigate the manipulation detection probability of EVM with 10 and 20 candidates in Figures 6.4 (a) and 6.4 (b), respectively. In each case, the numbers of vote changing attempts are realistically considered to be three times the number of candidates or more. Even for as few as three times N attempts, the probability of detection is as high as 97% and 96% for $N = 10$ and 20, respectively. Naturally, the probability of manipulation-detection increases with M and for $M \geq 6N$, the detection probability is almost 100% in both cases.

6.5.2 Preventing Manipulation by Dummy Votes

Without any control over the choice of anonymized subset by the voter, full de-anonymization may not be achieved from the casted AVs for some of the permutation groups. In such cases, the EVM would generate a set of dummy AVs that would lead to full de-anonymization. This is possible for a compromised EVM to generate the set of dummy AVs in such a way that together with the casted AVs they become consistent with a different candidate-order than the one displayed to the voters. Consequently, the vote counting would be totally misled.

This type of manipulation can be detected with the hash of the candidate-orders generated by the RNG which is supplied to the BB at the end of the vote. The hash of the candidate-orders, constructed after complete de-anonymization, would be computed. This computed hash would not match the one provided by the RNG if the set of dummy AVs are manipulated. Note that the expected requirement of dummy AVs is negligible when the

number of voters for each group exceeds even 20 times the number of candidates (see Section 6.6.2).

6.5.3 Mitigation of Threat on Privacy

Voter privacy has to be retained in two phases, i.e., before and after the receipt exchange. Before the receipt exchange, each voter has to validate her receipt by a polling officer. The standard procedure to validate a document is to use both digital mechanism (by digital signature, watermark, etc.) and manual on-the-spot verification. In this context, it ensures that a fake receipt would not be produced even when the receipt printing authority is compromised. It also prevents the clash attack as reported in [142] where a number of compromised voters would drop some fake receipts and take their own one with them so that these are never verified.

If there is a possibility to associate a voter with her receipt at any point, her privacy is violated. If the RNG is compromised, the voter can be identified from the receipt published in the BB and from the physical sequence of appearance of a voter. Thus, an untrusted EVM is not used in the proposed scheme to generate the receipt numbers. Instead, the proposed scheme uses the receipts printed by a trusted, simple, hardware-only external RNG.

For manual validation, the voter does not need to show the receipt explicitly to the polling officer; however, while signing the receipt the polling agent has enough opportunity to notice it. If the polling officer is compromised by a candidate and reveals the vote of a person, the voter's privacy is hampered and coercion or vote trading may occur. This is why a direct vote would not serve the purpose. However, since our proposed system generates AVs with k-anonymity of the candidates, the threat on privacy is mitigated at this level.

We also adopt the floating-receipt technique to prevent coercion or vote trading. After validation, the polling agent would drop each receipt in an exchange box and the voter will take another one at random. It will be ensured either mechanically (say, the exit door will not open until the receipt is exchanged) or by observation of all-party polling agents. Once this exchange is complete, there is no link between the voter and her receipt or vote. So, from then on, the threat on privacy is naturally mitigated as the coercer would not be able to have any information from the receipt a voter carries and, on top of it, the voter cannot prove the way she voted.

In the next section, we discuss some optimization issues related to the selection of dummy votes and how they impact the voting process.

6.6 Optimization Issues

An anonymized EVS needs to be completely random since any specific technique would allow the compromised authority to manipulate cast votes, or a potential coercer may reveal a vote following this pattern. As the random approach requires a significantly large number of votes to achieve full de-anonymization, we introduced the concept of dummy votes in Section 6.4.1. We proposed to construct a minimal required set of AVs ($S_{\min}$) for a given N and k in order to achieve full de-anonymization. Once the voting is done, this minimal set is matched against the collected votes and the missing votes required to achieve full de-anonymization are identified and appended as dummy votes. Dummy votes are usually required when the number of voters is very low, implying that not enough variety of AVs has been cast for full de-anonymization. $S_{\min}$ can be constructed in a number of ways for the same candidate-group. The careful selection of the most preferable $S_{\min}$ can effectively reduce the number of dummy votes required to bring full de-anonymization, ensuring less communication overhead.

6.6.1 Selection of S_{min} Using Pre-election Polling

The selection of an $S_{\min}$ that would reduce the requirements of dummy votes cannot be made optimal since there is no control over the choice of candidates in AVs. Hence, we propose a near-optimal heuristic for this selection using pre-polling survey results which are widely used for other purposes around the world. The performance of this approach is evaluated by simulation and will be presented in the next section.

By the laws of statistics, we can assume that any dramatic change from the pre-polling survey result is the least probable case, whereas the swapping of any two candidates of the lowest popularity gap is the most probable one. Keeping this in mind we also aim to establish that the less it deviates from the pre polling survey result, the less will be the required number of dummy votes. We aim to utilize the pre-polling survey result to minimize the required number of dummy votes to avoid the risk of manipulation without interpreting the actual votes cast.

Using the candidate ranking from the pre-election survey result, the occurrence of candidates in an $S_{\min}$ can be selected in order of their popularity. For example, for the

popularity order of candidates $[A\ B\ C\ D]$, following this heuristic $S_{\min}$ may consist of 7 AVs, $\{A,B,C\}:1,\ \{A,B,D\}:1,\ \ \{A,C,D\}:1,\ \{A,B,C\}:2,\ \{A,B,D\}:2,\ \{A,B,C\}:3,$ and $\{-,-,D\}:4$. It inherently assumes that A will have more AVs than any other one and so on. Again, if we have a pre-voting polling result in our hand implying that the popularity order for candidates is $[D\ B\ C\ A]$, we can construct the $S_{\min}$ as: $\{A,B,D\}:4$,$\{A,C,D\}:4$, $\{B,C,D\}:4$,$\{A,B,D\}:2,\ \{B,C,D\}:2,\ \ \{B,C,D\}:3,$ and $\{A,-,-\}:1$. On the contrary, if the opposite happens then we might end up generating dummy votes as high as 6 in the worst case scenario when the votes are cast in completely reverse order of the pre-poll result. However, we consider this to be highly unlikely.

In our analysis, we were interested to see if the forecast remains true while voting then whether the minimum number of dummy votes is required compared to other possible $S_{\min}$s. The significance of this validation is that now we can ensure that even anyone suspecting manipulation in dummy votes can be assured that such minimal manipulated votes can merely change the election result. For example, let us assume that in a particular EVM for a particular candidate-order group, $S_{\min}$ is constructed using the pre-polling survey result and after vote casting only 2 dummy votes were found to be required to achieve full de-anonymization. Now, if the winning candidate wins by more than 4 votes, then we can confirm that even dummy vote manipulation can make no alteration to the final result.

The number of voters for each candidate-group has a direct impact on the requirement of dummy vote for that group and, instead, all the voters are distributed among the groups. Hence, it is obvious that the decision of how many groups will be used is also significant in this respect. If a small number of groups are used, the need for dummy votes is expected to be low. However, as discussed earlier, the more groups are used, the better the situation from a privacy perspective. This trade-off is addressed by the empirical analysis presented below.

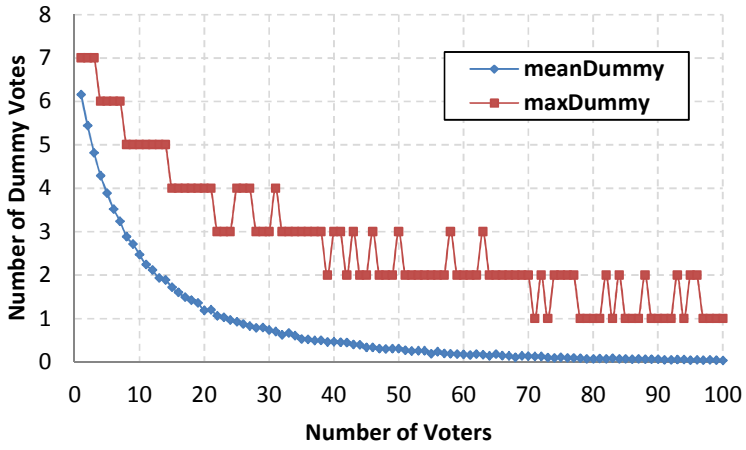


Figure 6.5: Required number of Dummy Vote trend for different number of voters.

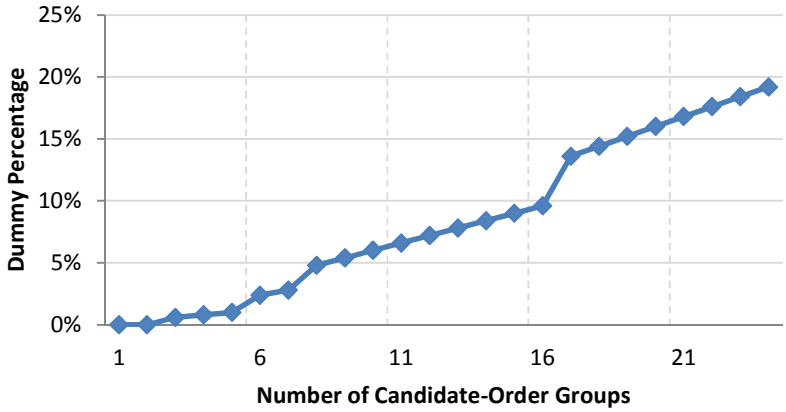


Figure 6.6: Required dummy vote percentage for different number of candidate groups.

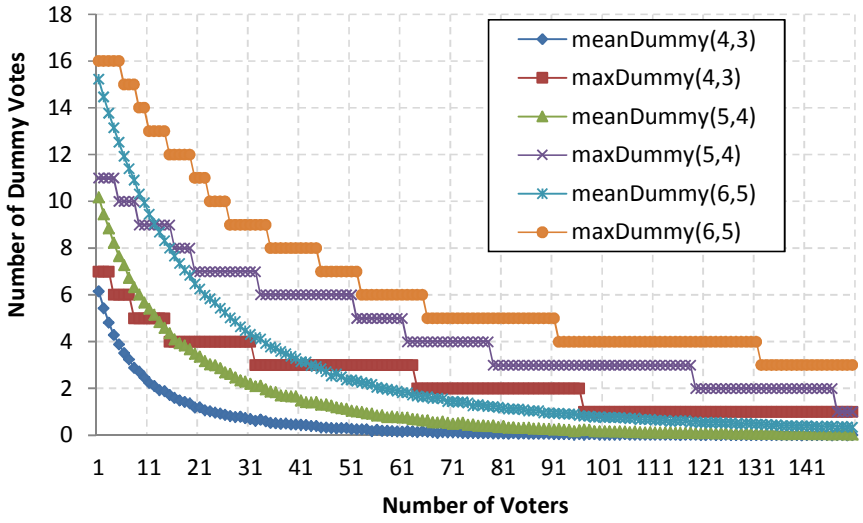
6.6.2 Empirical Analysis

In the simulation based empirical analysis, we have used the candidate-order group $[A\ B\ C\ D]$ with a pre-polling survey result showing the probability of popularity distribution as [0.05 0.25 0.1 0.6]. Figure 6.5 shows the typical trend signifying that the required number of dummy votes reduces as the number of voters is increased. Here, the mean dummy vote trend is shown to portray the average case, whereas the trend for the maximum dummy vote represents the worst case scenario.

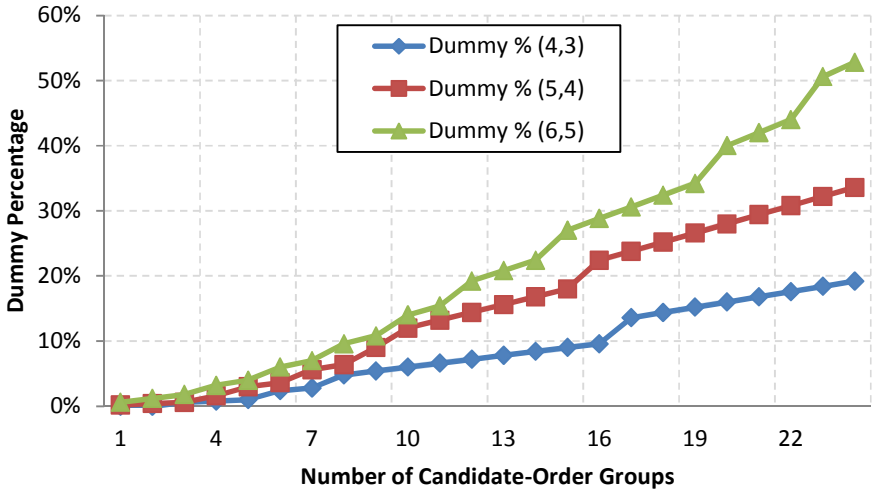
For this four candidate scenario, there can be total $N! = 24$, number of candidate-order groups. Typically, we can assume that the number of voters in each polling booth is 500. Figure 6.6 shows that the percentage of required dummy votes effectively increases as the number of total candidate-order groups used is increased. Using this approach, we can now design the voting system to accommodate user specific requirements. The higher the

number of candidate-order groups, the more robust will be the voting system from the security and privacy points of view. For a fixed number of total voters, in contrast, the higher the number of candidate-order groups, fewer the number of voters per group will be which in turn implies that more dummy votes will be required. We need to keep these both in mind while designing a voting system. For example, if it is mentioned that no more than

10% of the votes can be allowed as dummy votes, then from Figure 6.6, we can state that highest 16 total candidate-order groups can be allowed for maintaining the requirement. Then again, if a voting system targets to work with total 20 candidate-order groups, then it should also realize that it may have to bear dummy vote percentage as high as 16% to fulfill that criterion.



(a)



(b)

Figure 6.7: For candidate number $N = \{4,5,6\}$, (a) Dummy Vote trend for different number of voters and (b) Dummy Vote percentage for different number of candidate groups.

Table 6.3: Dummy Votes for All Possible Variations from the Pre Polling Survey Result with Popularity Order $[D\ B\ C\ A]$

Actual Popularity Order	Dummy (Mean)	Dummy (Max)
$[A\ C\ B\ D]$	1.646	5
$[A\ D\ B\ C]$	1.259	3
$[A\ B\ C\ D]$	2.156	4
$[A\ B\ D\ C]$	1.378	4
$[A\ C\ D\ B]$	1.24	4
$[A\ D\ C\ B]$	0.894	3
$[B\ C\ A\ D]$	1.434	3
$[B\ D\ A\ C]$	0.671	2
$[C\ B\ A\ D]$	2.015	3
$[D\ B\ A\ C]$	1.02	2
$[D\ C\ A\ B]$	0.913	2
$[C\ D\ A\ B]$	2.062	3
$[C\ A\ B\ D]$	1.602	4
$[D\ A\ B\ C]$	0.813	3
$[B\ A\ C\ D]$	2.309	3
$[B\ A\ D\ C]$	1.063	3
$[C\ A\ D\ B]$	0.767	3
$[D\ A\ C\ B]$	1.769	3
$[D\ C\ B\ A]$	0.911	3
$[C\ D\ B\ A]$	0.76	3
$[D\ B\ C\ A]$	0.252	2
$[C\ B\ D\ A]$	0.889	3
$[B\ C\ D\ A]$	0.888	2
$[B\ D\ C\ A]$	0.285	2

In Figure 6.7, we analysed the variation in these trends for varying number of candidates, $N = \{4,5,6\}$. For all the cases, the subsets used in vote anonymizing avail the highest possible level of anonymity, i.e., $k = N - 1$. The trends show that dummy vote requirement increases with the higher values of N.

So far, all the results have been produced assuming that the pre polling survey result is sustained in the actual vote. Now, let us focus on the possibilities of variation in the actual vote from the forecast and how the dummy votes are affected by this variation. A summary of this variation is presented in tabular format as shown in Table 6.3. We have assumed that

the pre-polling survey result in our hand implies that the popularity ranking for candidates is $[D\ B\ C\ A]$. For 1000 simulation runs, we found that the required dummy vote was lowest in the expected case when there was no variation from the survey result (highlighted in Table 6.3). Here, again the first column of dummy vote represents the average case whereas the latter depicts the worst case scenario.

6.7 Conclusion

In this chapter, we have proposed a k-anonymized electronic voting scheme that does not trust the EVM used in the voting process. The scheme successfully deals with the significant challenge of simultaneously protecting the privacy of voters and providing verifiability of the results, exploiting our novel subset-coding technique which is used for all the anonymization schemes presented in this thesis. The joint de-anonymization approach mitigates the threat of EVM level manipulation by virtue of its inherent consistency-preservation property and non-dependency of the candidate-order information used during anonymization. The additional use of standard cryptographic hashing and floating receipt concept ensure E2E verifiability while preventing coercion and vote trading.

Chapter 7

7 Conclusions and Future Works

The major challenge in research on privacy is that sometimes the quality and the verifiability of the data have to be preserved simultaneously. Since the obfuscation or precision-reduction based traditional approaches have limited success to achieve all the requirements satisfactorily, in recent times researchers have focused on the problem, especially in emerging areas of mass electronic communication. In this thesis, we have worked on solving this problem by developing novel techniques and presenting comprehensive relevant analyses. To establish our hypothesis, we have chosen two people-centric systems, i.e., PSS and EVS, as application scenario since together their characteristics cover the main aspects of such systems. Critical knowledge from the observations made in this research is not only valuable for our current domain, but also has a significant impact on relevant and related research fields.

In pursuing our aims, the key achievements of the thesis, along with their significance, are summarised as follows:

- The main innovation of this work is that the typical trade-off problem of maintaining privacy and data integrity at the same time is handled intelligently such that the desirable properties are maintained at their relevant points. Consequently, we have established that it is possible to achieve them simultaneously. The proposed scheme is applicable to emerging people centric systems that have a probability of "multiple observations of individual instances." Thus, the proposed scheme has prospective applicability to many multivariate systems.
- Replication is a well-established concept in data communication that inherently maintains trustworthiness/verifiability. For example, replication is done in network

communication to confirm acknowledgement, although it is not the optimal way. In our scheme, we have exploited the redundancy in replication to ensure trustworthiness/ verifiability. Our contribution in this regard is that we have devised the optimal replication pattern. In Chapter 5, we have aimed to anonymize observations in such a way that full decodability can be achieved with minimal observations, which in a sense minimally exploits this replication redundancy. Additionally, the concept of a minimum anonymized set of votes that may achieve full decodability is established in Chapter 6. In both cases, we have found that it is not necessary to obtain multiple observations of all the individuals to obtain full decodability.

- To obtain full and accurate information, we have devised a framework for joint decoding of some seemingly independent observations. To deal with multiple observations of the same object is a straight forward problem, whereas in our scheme we have dealt with multiple apparently autonomous observations that are occurring multiple times. Thus, we have jointly achieved trustworthiness through minimal replication redundancy and accurate POI-attribute association using our unified platform. Hence, we can not only provide data integrity with a reasonably low number of observations, but also use it to further enhance protection against untrustworthy behaviour, as well as to provide some kind of verifiability assurance. Our approach can be applied in domains where optimization is required in multiple dimensions.
- Our proposed scheme of simultaneously achieving privacy and data integrity indirectly suggests how the participation rate in public networks can be influenced. In the people-centric application scenarios presented in this thesis, many volunteers are required to participate in making the service available. Even in the absence of any reward scheme, the service provided itself can become sufficient to motivate volunteers to participate, if the fulfilment of their other demands can be assured. With the minimization of cost per service by optimizing required data size, and also the improvement of service quality with maintenance of data accuracy, the participation of the volunteers can be influenced positively in such systems. At the same time, in EVS, the inherent trustworthiness or verifiability with privacy assurance can boost voters' participation in cases where voting is not made compulsory.
- There have been very few approaches proposed so far that have dealt with the privacy data integrity trade-off issue, let alone addressing all other such sensitive issues. As

proposed in our scheme, all the desirable sensitive concerns like trust, verifiability, privacy, and data integrity of a people-centric system can be dealt by the same framework. From this perspective, this research work has made a notable contribution.

- Risk analysis and possible adversary modelling has also been comprehensively accomplished in this thesis. We have established that risk involved can be attenuated to remain within the user-defined threshold. We have identified and analysed the impact of the adjustable parameters of a people-centric application scenario. Note that in such systems some parameters are user controllable, some are naturally controlled from the environment, and some are dominated by system approach and behaviour. Thus, we have devised a possible way to achieve certain characteristics or property, as shown in Chapter 3. For example, we have shown that attenuating a parameter as simple as the number of friends, can keep the risk within a desired upper bound.

While addressing each of the issues mentioned above, extensive experiments have been performed to evaluate the proposed mechanisms. The research findings presented in the thesis can be extended to the following areas:

- The proposed subset coding scheme can be extended to deal with multidimensional privacy. For example, in a consumer price sharing application scenario, a user may not care about the presence in a super store, but will be very sensitive about the purchase of a particular item such as alcohol or a particular drug. In this case, product-level privacy may be incorporated with spatial privacy. Moreover, adversary inference often depends heavily on the temporal correlation of spatial occurrence. Hence, incorporating these additional dimensions to the location privacy scheme poses other research issues.
- Studies on socio-cultural behaviour suggest that different people from different countries react differently to privacy issues and also to a suspected privacy breach. Hence, the proposed subset coding scheme can be upgraded to provide different options of desired anonymity levels to work with. It can be made more attractive to users by allowing them to set the parameters and relative priority for privacy or accuracy.
- To encourage more voters to participate, critical knowledge gathered from this research work can be used to establish an analytical model of an EVS which can totally avoid the trusted hardware assumption.

- Preferential voting is practised in many countries around the world. It is, thus, worth exploring the applicability of our proposed privacy-integrity management scheme using *subset-coding* and joint de-anonymization in the process of preferential voting.
- The proposed scheme can be enhanced for real environments by incorporating various realistic incentive/reward conditions, to ensure sufficient participation in the system.
- Another challenging research problem is to include a reputation-based scheme for incorporating provision of outlier detection in our proposed system design.
- It is also worthwhile to explore our proposed scheme in other relevant application scenarios like online surveys, anonymous student feedback collection, anonymous unique incidence tracking and reporting [6] to achieve inferable privacy preservation and trustworthiness simultaneously.

People-centric applications have positive prospects with the advent of next generation wireless network and the Internet of Things. We are hopeful that the specific protocols and relevant results presented in this thesis will eventually improve the overall efficiency of achieving privacy and integrity at their respective desired points. Incorporation of this approach would increase the popularity of these systems to a great extent and would eventually make the services more meaningful and powerful. We hope that the critical knowledge obtained from this thesis will be considered in taking the first step to assure future cyber privacy and security so that people can enjoy every benefit of digital communication.

Bibliography

[1] A. Tanenbaum, Modern Operating Systems, Prentice-Hall: Upper Saddle River, NJ, USA, 1992.

[2] Y. Dong, S. Kanhere, C. Chou, and N. Bulusu, "Automatic collection of fuel prices from a network of mobile cameras," in Proceedings of the IEEE international Conference on Distributed Computing in Sensor Systems (DCOSS), 2008.

[3] S. Sehgal, S. Kanhere, and C. Chou, "Mobishop: using mobile phones for sharing consumer pricing information," in Demo Session of the IEEE international Conference on Distributed Computing in Sensor Systems (DCOSS), 2008.

[4] L. Deng and L. Cox, "Livecompare: grocery bargain hunting through participatory sensing," in Proceedings of the ACM workshop on Mobile Computing Systems and Applications, 2009.

[5] N. Bulusu, C. Chou, S. Kanhere, Y. Dong, S. Sehgal, D. Sullivan, and L. Blazeski, "Participatory sensing in commerce: using mobile camera phones to track market price dispersion," in Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys), 2008.

[6] J. Ballesteros, M. Rahman, B. Carbunar, and N. Rishe, "Safe cities. A participatory sensing approach," in Proceedings of the IEEE international Conference on Local Computer Networks (LCN), 2012.

[7] S. Eisenman, E. Miluzzo, N. Lane, R. Peterson, G. Ahn, and A. Campbell, "The Bikenet mobile sensing system for cyclist experience mapping," in Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys), 2007.

[8] B. Hull, V. Bychkovsky, and Y. Zhang, "CarTel: a distributed mobile sensor computing system," in Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys), 2006.

[9] R. Ganti, N. Pham, H. Ahmadi, S. Nangia, T. Abdelzaher, " GreenGPS: a participatory sensing fuel-efficient maps application," in Proceedings of the ACM Conference on Mobile Systems, applications, and services (MobiSys), 2010.

[10] P. Mohan, V. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys), 2008.

[11] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estin, and M. Hansen, "Image browsing, processing and clustering for participatory sensing: lessons from a DietSense prototype," in Proceedings of the Workshop on Embedded Networked Sensors (EmNetS), 2007.

[12] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, P. Boda, "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research," in Proceedings of the ACM international Conference on Mobile Systems, Applications, and Services (MobiSys), 2009.

[13] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, and R. Peterson, "People-centric urban sensing," in Proceedings of the international Wireless Internet Conference (WICON), 2006.

[14] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "SoundSense: scalable sound sensing for people-centric applications on mobile phones," in Proceedings of the ACM international Conference on Mobile Systems, Applications, and Services (MobiSys), 2009.

[15] "Quake-catcher network," Stanford University (http://qcn.stanford.edu), 2010.

[16] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe, "ParkNet: drive-by sensing of road-side parking statistics," in Proceedings of the ACM international Conference on Mobile Systems, Applications, and Services (MobiSys), 2010.

[17] F. Jazizadeh and B. Gerber, "Towards adaptive comfort management in office buildings using participatory sensing for end user driven control," in Proceedings of the ACM Workshop on Embedded Systems for Energy-Efficiency in Buildings (BuildSys), 2012.

[18] P. Zhou, Y. Zheng, and M. Li, "How long to wait? Predicting bus arrival time with mobile phone based participatory sensing," in Proceedings of the ACM international Conference on Mobile Systems, Applications, and Services (MobiSys), 2012.

[19] V. Vehovar and K. L. Manfreda, "Overview: online surveys," The SAGE Handbook of Online Research Methods, 2008.

[20] "Franken's location-privacy bill would close mobile-tracking 'loopholes'," http://www.wired.com/epicenter/2011/06/franken-location-loopholes/, accessed on November 29, 2011.

[21] D. Christin, A. Reinhardt, S. Kanhere, and M. Hollic, "A survey on privacy in mobile participatory sensing applications," Journal of Systems and Software, 84(11), 2011.

[22] D. Dill and D. Castro, "The US should ban paperless electronic voting machines," Communications of the ACM, 51(10), 2008.

[23] D. Chaum, A. Essex, R. Carback, J. Clark, S. Popoveniuc, A. T. Sherman, and P. L. Vora, "Scantegrity: end-to-end voter-verifiable optical-scan voting," IEEE Journal of Security & Privacy, vol. 6(3), pp. 40-46, 2008.

[24] D. Chaum, R. Carback, J. Clark, A. Essex, S. Popoveniuc, R. L. Rivest, P. Y. A. Ryan, E. Shen, A. T. Sherman, and P. L. Vora, "Scantegrity II: end-to-end verifiability by voters of optical scan elections through confirmation codes," IEEE Transaction on Information Forensics and Security, vol. 4, pp. 611–627, 2009.

[25] P. Y. A. Ryan, D. Bismark, J. Heather, S. Schneider, and Z. Xia, "Prêt à Voter: a voter-verifiable voting system," IEEE Transaction on Information Forensics and Security, vol. 4, pp. 662-673, 2009.

[26] T. Moran and M. Naor, "Split-Ballot Voting: everlasting privacy with distributed trust," in Proceedings of the ACM international Conference on Computer and Communications Security (CCS), 2007.

[27] R. L. Rivest and W. D. Smith, "Three voting protocols: ThreeBallot, VAV, and Twin," in Proceedings of the USENIX/ACCURATE Electronic Voting Technology Workshop (EVT), 2007.

[28] A. Essex and U. Hengartner, "Hover: trustworthy elections with hash-only verification," IEEE Security & Privacy, vol. 10(5), pp. 18-24, 2012.

[29] I. Rouf, R. Miller, H. Mustafa, T. Taylor, S. Oh, W. Xu, M. Gruteser, W. Trappe, and I. Seskar, "Security and privacy vulnerabilities of in-car wireless networks: a tire pressure monitoring system case study," in Proceedings of the USENIX Conference on Security, 2010.

[30] L. Hu and C. Shahabi, "Privacy assurance in mobile sensing networoks: go beyond trusted servers," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2010.

[31] T. Sabrina, and M. Murshed, "Privacy in participatory sensing systems," In J. Abawajy, M. Pathan, M. Rahman, A. Pathan, & M. Deris (Eds.), Network and Traffic Engineering in Emerging Distributed Computing Applications (pp. 124-143).

[32] T. Sabrina and M. Murshed, "Analysis of location privacy risk in a plain-text communication based participatory sensing system using subset coding and mix network," in Proceedings of the International Symposium on Communications and Information Technologies (ISCIT), 2012.

[33] M. Murshed, T. Sabrina, A. Iqbal, and K. H. Alam, "A novel anonymization technique to trade-off location privacy and data integrity in participatory sensing systems," in Proceedings of the IEEE international Conference on Network and System Security (NSS), 2010.

[34] M. Murshed, A. Iqbal, T. Sabrina, and K. H. Alam, "A subset coding based *k*-anonymization technique to trade-off location privacy and data integrity in participatory sensing systems," in Proceedings of the IEEE international Symposium on Network Computing and Applications (NCA), 2011.

[35] M. Murshed, T. Sabrina, A. Iqbal, and M. Ali, "Verifiable and privacy preserving electronic voting with untrusted machines," in Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2013.

[36] K. Huang, S. Kanhere, and W. Lu, "Preserving privacy in participatory sensing systems," Journal of Computer Communications, vol. 33(11), pp. 1266–1280, 2010.

[37] S. Russell and P. Norvig, "Artificial intelligence: a modern approach,"(2nd ed.), Prentice Hall, 2003.

[38] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. Srivastava, "Participatory sensing," in Proceedings of the 1st Workshop on World-Sensor-Web (WSW), 2006.

[39] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson, "People-centric urban sensing," in Proceedings of the International Wireless Internet Conference (WICON), 2006.

[40] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn, " The rise of people-centric sensing ," IEEE Internet Computing, vol. 12(4), pp. 12-21, 2008.

[41] A. Kapadia, D. Kotz, and N. Triandopoulos, "Opportunistic sensing: security challenges for the new paradigm," in Proceedings of the 1st international Conference on Communication Systems and Networks (COMNETS), 2009.

[42] G .Lee, W. Kim, and D. K. Kim, "An effective method for location privacy in ubiquitous computing," in Proceedings of the Embedded and Ubiquitous Computing, Springer Berlin Heidelberg, 2005.

[43] C. Wang and W. Ku, "Anonymous sensory data collection approach for mobile participatory sensing," in Proceedings of the 28th International Conference on Data Engineering Workshops (ICDE), 2012.

[44] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in Proceedings of the USENIX Security Symposium, 2004.

[45] (2012) Tor metrics - number of users. [Online]. Available: https://metrics.torproject.org/users.html.

[46] (2012) Tor metrics - number of relays. [Online]. Available: https://metrics.torproject.org/network.html.

[47] K. Bauer, J. Juen, N. Borisov, D. Grunwald, D. Sicker, and D. Mccoy, "On the optimal path length for Tor," in Proceedings of the Hot Topics in Privacy Enhancing Technologies Symposium (HotPETS), 2010.

[48] T.-W. Ngan, R. Dingledine, and D. S. Wallach, "Building incentives into Tor," Lecture Notes in Computer Science Volume 6052, pp 238-256, 2010.

[49] R. Dingledine and N. Mathewson. (2012) Tor path specification. [Online]. Available: https://git.torproject.org/checkout/tor/master/doc/spec/path-spec.txt.

[50] K. Bauer, D. McCoy, D. Grunwald, T. Kohno, and D. Sicker, "Low-resource routing attacks against Tor," in Proceedings of the ACM Workshop on Privacy in Electronic Society (WPES), 2007.

[51] R. Snader and N. Borisov, "A tune up for Tor: improving security and performance in the Tor network," in Proceedings of the Network and Distributed System Security Symposium (NDSS), 2008.

[52] M. Sherr, M. Blaze, and B. T. Loo, "Scalable link-based relay selection for anonymous routing," in Proceedings of the Privacy Enhancing Technologies Symposium (PETS), 2009.

[53] M. Akhoondi, C. Yu, and H. V. Madhyastha, "LASTor: a low-latency AS-aware Tor client," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2012.

[54] M. AlSabah, K. Bauer, T. Elahi, and I. Goldberg, "The path less travelled: overcoming Tor's bottlenecks with multipaths," University of Waterloo, Tech. Rep., 2011.

[55] H. Hsiao, T. Kim, A. Perrig, A. Yamada, S. Nelson, M. Gruteser, and W. Meng, "LAP: lightweight anonymity and privacy," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2012.

[56] A. Beresford and F. Stajano, "Location privacy in pervasive computing," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2003.

[57] K. Shilton, "Four billion little brother?: privacy, mobile phones, and ubiquitous data collection," Communications of the ACM, 52 (11), 2009.

[58] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in Proceedings of the ACM SIGMOD international conference on Management of data, 2005.

[59] K. Shilton, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Participatory privacy in urban sensing," in Proceedings of the International Workshop on Mobile Device and Urban Sensing (MODUS), 2008.

[60] B. Palanisamy, and L. Liu, "Mobimix: protecting location privacy with mix-zones over road networks," in Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2011.

[61] G. Zhong, and U. Hengartner, "A distributed *k*-anonymity protocol for location privacym," in Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), 2009.

[62] S. Gao, J. Ma, W. Shi, G. Zhan, and C.Sun, "TrPF: a trajectory privacy-preserving framework for participatory sensing," IEEE Transaction on Information Forensics and Security, vol. 8 (6), pp. 874-887, 2013.

[63] B. Gedik, and L. Liu, "Location privacy in mobile systems: a personalized anonymization model," in Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS), 2005.

[64] M. Gruteser, and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in Proceedings of the ACM international Conference on Mobile Systems, Applications, and Services (MobiSys), 2003.

[65] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in Proceedings of the ACM conference on Computer and Communications Security (CCS), 2007.

[66] J. Meyerowitz and R. Choudhury, "Hiding stars with fireworks: location privacy through camouflage," in Proceedings of the ACM international Conference on Mobile computing and networking (MobiCom), 2009.

[67] M. Duckham, and L. Kulik, "Simulation of obfuscation and negotiation for location privacy," in Spatial Information Theory, pp. 31-48. Springer Berlin Heidelberg, 2005.

[68] C. Ardagna, M. Cremonini, Er. Damiani, S. Vimercati, and P. Samarati, "Location privacy protection through obfuscation-based techniques," in Data and Applications Security XXI, pp. 47-60. Springer Berlin Heidelberg, 2007.

[69] M. F. Mokbel, C. Chow, and Walid G. Aref. "The new Casper: query processing for location services without compromising privacy," in Proceedings of the international Conference on Very Large Data Bases (VLDB), 2006.

[70] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries," IEEE Transactions on Knowledge and Data Engineering, vol. 19(12), pp. 1719-1733, 2007.

[71] T. Xu, and Y. Cai, "Location cloaking for safety protection of ad hoc networks," in Proceedings of the IEEE international Conference on Computer Communications (INFOCOM), 2009.

[72] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, and D. Kotz, "AnonySense: opportunistic and privacy-preserving context collection," in Proceedings of the Pervasive Computing and Communications (PerCom), 2008.

[73] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*l*-diversity: privacy beyond *k*-anonymity," ACM Transactions on Knowledge Discovery from Data, vol 1(1), article 3, 2007.

[74] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in Proceedings of the ACM international Conference on World Wide Web (WWW), 2008.

[75] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu, "Spacetwist: managing the trade-offs among location privacy, query performance, and query accuracy in mobile services," in Proceedings of the IEEE international Conference on Data Engineering (ICDE), 2008.

[76] K. L. Huang, S. S. Kanhere, and W. Hu, "Towards privacy-sensitive participatory sensing," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2009.

[77] C.-Y. Chow, M. F. Mokbel, and X. Liu, "A peer-to-peer spatial cloaking algorithm for anonymous location-based service," in Proceedings of the ACM international Symposium on Advances in Geographic Information Systems, 2006.

[78] T. Hashem, and L. Kulik, "Safeguarding location privacy in wireless ad-hoc networks," in Proceedings of the In Ubicomp 2007: Ubiquitous Computing, pp. 372-390. Springer Berlin Heidelberg, 2007.

[79] I. Rodhe, C. Rohner, and E. C-H. Ngai, "On location privacy and quality of information in participatory sensing," in Proceedings of the ACM Symposium on QoS and Security for Wireless and Mobile Networks, 2012.

[80] T. Hashem, L. Kulik, and R. Zhang, "Privacy preserving group nearest neighbor queries," in Proceedings of the ACM international Conference on Extending Database Technology, 2010.

[81] K. Vu, R. Zheng, and J. Gao, "Efficient algorithms for *k*-anonymous location privacy in participatory sensing," in Proceedings of the IEEE international Conference on Computer Communications (INFOCOM), 2012.

[82] H. Takabi, J. BD Joshi, and H. A. Karimi, "A collaborative *k*-anonymity approach for location privacy in location-based services," in Proceedings of the IEEE international Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2009.

[83] E De Cristofaro and C. Soriente, "Participatory privacy: enabling privacy in participatory sensing," Network, IEEE 27, no. 1 (2013): 32-36.

[84] C. Blundo, A. De Santis, G. Di Crescenzo, A. G. Gaggia, and U. Vaccaro, "Multi-secret sharing schemes," in Advances in Cryptology — CRYPTO '94, pp 150-163, doi-10.1007/3-540-48658-5_17, vol. 839.

[85] F. Durr, P. Skvortsov, and K. Rothermel, "Position sharing for location privacy in non-trusted systems," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2011.

[86] G. F. Marias, C. Delakouridis, L. Kazatzopoulos, and P. Georgiadis, "Location privacy through secret sharing techniques," in Proceedings of the IEEE international Symposium on World of Wireless Mobile and Multimedia Networks (WoWMoM), 2005.

[87] M. Wernke, F. Durr, and K. Rothermel, "PShare: position sharing for location privacy based on multi-secret sharing," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2012.

[88] S. Mascetti, D. Freni, C. Bettini, X. S. Wang, and S. Jajodia, "Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies," Journal on VLDB, vol. 20(4), pp. 541-566, 2011.

[89] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: anonymizers are not necessary," in Proceedings of the ACM SIGMOD international Conference on Management of data, 2008.

[90] E. De Cristofaro, A. Durussel, and I. Aad, "Reclaiming privacy for smartphone applications," in Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), 2011.

[91] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in Proceedings of the International Conference on Pervasive Services (ICPS), 2005.

[92] P. Shankar, V. Ganapathy, and L. Iftode, "Privately querying location-based services with sybilquery," in Proceedings of the ACM international Conference on Ubiquitous Computing (UbiComp), 2009.

[93] P. Skvortsov, F. Dürr, and K. Rothermel, "Map-aware position sharing for location privacy in non-trusted systems," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2012.

[94] C. W. Chan and C. C. Chang, "A scheme for threshold multi-secret sharing," Applied Mathematics and Computation, vol 166(1) 2005.

[95] I. Boutsis and V. Kalogeraki, "Privacy preservation for participatory sensing data," in Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), 2013.

[96] D. Christin, D. R. P.-Sorolla, S. S. Kanhere, and M. Hollick, "TrustMeter: a trust assessment framework for collaborative path hiding in participatory sensing applications," Technical Report TR-SEEMOO-2012-02.

[97] D. Christin, J. Guillemet, A. Reinhardt, M. Hollick, and S. S. Kanhere, "Privacy-preserving collaborative path hiding for participatory sensing applications," in Proceedings of the IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), 2011.

[98] S. Reddy, V.t Samanta, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "MobiSense—mobile network services for coordinated Participatory Sensing," in Proceedings of the International Symposium on Autonomous Decentralized Systems (ISADS), 2009.

[99] J. Krumm, "Inference attacks on location tracks," in Proceedings of the international Conference on Pervasive Computing, 2007.

[100] R. K, Ganti, N. Pham, Y.-E. Tsai, and T. F. Abdelzaher, "PoolView: stream privacy for grassroots participatory sensing," in Proceedings of the ACM conference on Embedded network Sensor Systems (SenSys), 2008.

[101] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, and T. Abdelzaher, "Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing," Wireless Sensor Networks, Springer Berlin Heidelberg, 2010.

[102] J. Shi, R. Zhang, Y. Liu, and Y. Zhang, "Prisense: privacy-preserving data aggregation in people-centric urban sensing systems," in Proceedings of the IEEE international Conference on Computer Communications (INFOCOM), 2010.

[103] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. T. Abdelzaher, "Pda: Privacy-preserving data aggregation in wireless sensor networks," in Proceedings of the IEEE international Conference on Computer Communications (INFOCOM), 2007.

[104] L. Sweeney, "*k*-anonymity: a model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 2002.

[105] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information ," In ACM PODS, 1996.

[106] W. Winkler, "Using simulated annealing for k-anonymity," Research Report 2002-07, US Census Bureau Statistical Research Division, 2002.

[107] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2005.

[108] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Blocking anonymity threats raised by frequent itemset mining," in Proceedings of the IEEE International Conference on Data Mining, 2005.

[109] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "*k*-anonymous patterns," In Knowledge Discovery in Databases (PKDD), Springer Berlin Heidelberg, 2005.

[110] A. Beach, M. Gartrell, and R. Han, "Social-*k*: real-time *k*-anonymity guarantees for social network applications," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications Workshops (PerCom), 2010.

[111] H. Choi, S. Chakraborty, Z. M. Charbiwala, and M. B. Srivastava, "Sensorsafe: a framework for privacy-preserving management of personal sensory information," In Secure Data Management, pp. 85-100. Springer Berlin Heidelberg, 2011.

[112] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2012.

[113] E. De Cristofaro and R. Pietro, "Adversaries and countermeasures in privacy-enhanced urban sensing systems," IEEE Systems Journal, 2012.

[114] E. De Cristofaro and C. Soriente, "PEPSI: Privacy-enhanced participatory sensing infrastructure," in Proceedings of the ACM conference on Wireless network security (WiSec), 2011.

[115] M. Gadzheva, "Location privacy in a ubiquitous computing society," International Journal of Electronic Business, vol. 6(5), pp. 450-461, 2008.

[116] D. Christin, and M. Hollick, "Roadmap for privacy protection in mobile sensing applications," In European Data Protection: Coming of Age, Springer Netherlands, 2013.

[117] A. Ruzzelli, R. Jurdak, and G. OŠHare, "Managing mobile-based participatory sensing communities," in Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SENSYS), 2007.

[118] X. Xie, H. Chen, and H. Wu, "Bargain-based stimulation mechanism for selfish mobile nodes in participatory sensing network," in Proceedings of the IEEE international Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2009.

[119] J.-S. Lee and Baik Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," in Proceedings of the IEEE international Conference on Pervasive Computing and Communications (PerCom), 2010.

[120] G. Danezis, S. Lewis, and R. J. Anderson, "How much is location privacy worth?," in Proceedings of the annual Workshop on the Economics of Information Security (WEIS), 2005.

[121] J.-S. Lee and B. Hoh, "Dynamic pricing incentive for participatory sensing," in Proceedings of the IEEE international Conference on Pervasive and Mobile Computing, 2010.

[122] A. Dua, N. Bulusu, W.-C. Feng, and W. Hu, "Towards trustworthy participatory sensing," in Proceedings of the USENIX Workshop on Hot Topics in Security (HotSec), 2009.

[123] S. Saroiu and A. Wolman, "I am a sensor, and I approve this message," in Proceedings of the ACM Workshop on Mobile Computing Systems & Applications, 2010.

[124] K. L. Huang, S. S. Kanhere, and W. Hu, "Are you contributing trustworthy data?: the case for a reputation system in participatory sensing," in Proceedings of the ACM international Conference on Modeling, analysis, and simulation of wireless and mobile systems, 2010.

[125] J. Epstein, "How things work: electronic voting," IEEE Computers, 2007.

[126] C. Paar and J. Pelzl, "Hash functions", in Understanding cryptography, a textbook for students and practitioners, Springer, 2009.

[127] D. Bernhard, V. Cortier, O. Pereira, and B. Warinschi, "Measuring vote privacy, revisited," in Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2012.

[128] J. Dreier, P. Lafourcade, and Y. Lakhnech, "A formal taxonomy of privacy in voting protocols," in Proceedings of the IEEE Workshop on Security and Forensics in Communication Systems (SFCS), 2012.

[129] T. P. Pedersen, "Non-interactive and Information-Theoretic Secure Verifiable Secret Sharing," in Proceedings of the Advances in Cryptology (CRYPTO), 1991.

[130] J. Bohli, J. Quade, and S. Röhrich, "Bingo voting: secure and coercion-free voting using a trusted random number generator," Lecture Notes in Computer Science, 4896, pp. 111-124, 2007.

[131] J-M. Bohli, C. Henrich, C. Kempka, J. M-Quade, and S. Röhrich, "Enhancing electronic voting machines on the example of bingo voting," IEEE Transaction on Information Forensics and Security, 4(4), 2009.

[132] A. Essex, J. Clark, U. Hengartner, and C. Adams, "Eperio: mitigating technical complexity in cryptographic election verification," IACR Cryptology ePrint Archive, 2012, 178.

[133] A. Essex and U. Hengartner, "Oblivious printing of secret messages in a multi-party setting," in Proceedings of the international Conference on Financial Cryptography and Data Security (FC 12), 2012.

[134] A. Essex, C. Henrich, and U. Hengartner, "Single layer optical-scan voting with fully distributed trust," in Proceedings of the international Conference on E-voting and Identity (VoteID), 2011.

[135] S. Popoveniuc and B. Hosp, "An introduction to punchscan. Threat analyses for voting system categories," in Proceedings of the Workshop on Rating Voting Methods (VSRW), 2006.

[136] S. Popoveniuc and B. Hosp, "An introduction to punchscan," in Proceedings of the IAVoSS Workshop On Trustworthy Elections (WOTE), 2006.

[137] C. A. Neff, "Practical high certainty intent verification for encrypted votes," VoteHere, 2004.

[138] B. Adida and C. A. Neff, "Ballot casting assurance," in Proceedings of the USENIX/ACCURATE Electronic Voting Technology Workshop (EVT), 2006.

[139] T. Moran and M. Naor, "Receipt-free universally-verifiable voting with everlasting privacy," in Proceedings of the Advances in Cryptology (CRYPTO), 2006.

[140] R. G. Costa, A. O. Santin, and C. A. Maziero, "A three-ballot-based secure electronic voting system," in Proceedings of the IEEE international Conference on Security and Privacy (S&P), 2008.

[141] R. Kusters, T. Truderung, and A. Vogt, "Verifiability, privacy, and coercion-Resistance: new insights from a case study," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2011.

[142] R. Kusters, T. Truderung, and A. Vogt, "Clash attacks on the verifiability of e-voting systems," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2012.

[143] K. Henry, D. R. Stinson, and J. Sui, "The effectiveness of receipt-based attacks on ThreeBallot," IEEE Transaction on Information Forensics and Security, 4(4), 2009.

[144] Z. Xia, S.A. Schneider, J. Heather, P.Y.A. Ryan, D. Lundin, R. Peel, P. Howard, "Prêt à Voter: all-in-one," in Proceedings of the IAVoSS Workshop On Trustworthy Elections (WOTE), 2007.

[145] P. Ryan, T. Peacock, "A threat analysis of Prêt à Voter ," Towards Trustworthy Elections, Springer-Verlag Berlin, 2010.

[146] Z. Xia, C. Culnane, J. Heather, H. Jonker, P.Y.A. Ryan, S. Schneider, S. Srinivasan, "Versatile Prêt à Voter: handling multiple election methods with a unified interface," in Proceedings of the International Conference on Cryptology in India (INDOCRYPT), 2010.

[147] J. Graaf, "Voting with unconditional privacy by merging Prêt à Voter and PunchScan," IEEE Transaction on Information Forensics and Security, 4(4), 2009.

[148] D. Demirel, M. Henning, J. graaf, P. Ryan, and J. Buchmann, "Prêt à Voter providing everlasting privacy," in Proceedings of the E-Voting and Identify (VoteID), LNCS 2013.

[149] B. Adida and R. L. Rivest, "Scratch & vote: self-contained paper-based cryptographic voting ," in Proceedings of the ACM Workshop on WPES, 2006.

[150] R. Ara´ujo, R. Cust´odio, J. Van de Graaf, "A verifiable voting protocol based on Farnel. In: Chaum, D., Jakobsson, M., Rivest, R.L., Ryan, P.Y.A., Benaloh, J., Kutylowski,M., Adida, B. (eds.) Towards Trustworthy Elections. LNCS, Springer, Heidelberg, 2010.

[151] R. Araujo, P. Ryan, "Improving the Farnel voting scheme," Electronic Voting, 2008.

[152] M. R. Clarkson, S. Chong, and A. C. Myers, "Civitas: Toward a Secure Voting System," in Proceedings of the IEEE Symp. Security and Privacy (S&P), 2008.

[153] M. Raykova and D. Wagner, "Verifiable Remote Voting with Large Scale Coercion Resistance," Tech. Rep. CUCS-041-11, 2011.

[154] X. Yi and E. Okamoto, "Practical Internet Voting System," IEEE Transactions on Network and Computer Applications, 36(4) 2013.

[155] G. Schryen, E. Rich, E., "Security in Large-Scale Internet Elections: A Retrospective Analysis of Elections in Estonia, The Netherlands, and Switzerland, " IEEE Transactions on Information Forensics and Security, 4(4), 2009.

[156] T. Carrol and D. Grosu, "A Secure and Anonymous Voter-Controlled Election Scheme," Journal of Computer and network Applications, 2009 .

[157] C. Fan and W. Sun, “An Efficient Multi-Receipt Mechanism for Uncoercible Anonymous Electronic Voting,” International Journal on Mathematical and Computer Modelling, 48(5), 2008.

[158] S. Davtyan, A. Kiayias, L. Michel, A. Russel, and A. A. Shvartsman, “Integrity of electronic voting systems: fallacious use of cryptography,” in Proceedings of the ACM Symposium on Applied Computing, 2012.

[159] D. Chaum, M. Jakobsson, R. Rivest, P. Ryan, J. Benaloh, M. Kutylowski, B. Adida (Eds.): Towards Trustworthy Elections, New Directions in Electronic Voting. Lecture Notes in Computer Science, Springer, 2010.

[160] R. Rivest, "Security of Voting Systems," in Proceedings of the USENIX Symposium on Networked System Design and Implementation (NSDI), 2007

[161] L. Langer, H. Jonker, and W. Pieters, “Anonymity and Verifiability in Voting: Understanding (Un)Linkability,” in Proceedings of the ACM International Conference on Information and Communications Security (ICICS), 2007.

[162] D. J. Reynolds, "A method for electronic voting with Coercion-free receipt," in Proceedings of the Frontiers in Electronic Elections, (FEE), 2005.

[163] D. Chaum, "E-voting: Secret-ballot receipts: True voter-verifiable elections," in Proceedings of the IEEE Security and Privacy (S&P), 2004.

[164] P. Ryan, "A variant of the Chaum voter-verifiable scheme," in Proceedings of the Annual Workshop on Information Technologies & Systems (WITS), 2005.

[165] T. Moran and M. Naor, "Receipt-free universally-verifiable voting with everlasting privacy," in Proceedings of the International Cryptology Conference (CRYPTO), 2006.

[166] J. D. Cohen(Benaloh) and M. J. Fischer "A robust veri_able cryptographically secure election scheme," in Proceedings of the Symposium on Foundations of Computer Science (FOCS), 1985.

[167] R. Cramer, M. Franklin, B. Schoenmakers, and M. Yung, "Multi-authority secret-ballot elections with linear work," in Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt), 1996.

[168] R. Cramer, R. Gennaro, and B. Schoenmakers, "A secure and optimally e_cient multi-authority election scheme," in Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt), 1997.

[169] A. Fujioka, T. Okamoto, and K. Ohta, "A practical secret voting scheme for large scale elections," in Proceedings of the Advances in Cryptology (AUSCRYPT), 1992.

[170] M. Hirt and K. Sako, " Efficient receipt-free voting based on homomorphic encryption," in Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt), 2000.

[171] J. Benaloh and D. Tuinstra, "Receipt-free secret-ballot elections," in Proceedings of the ACM Symposium on Theory of Computing (STOC), 1994.

==========